PHYSICAL SCIENCES-II
(Physics & Chemistry)
(w.e.f 2019-20)

Second Year

Intermediate Vocational

Bridge Course

STATE INSTITUTE OF VOCATIONAL DUCATION

BOARD OF INTERMEDIATE EDUCATION, A.P.

Smt. B.UDAYA LAKSHMI, I.A.S.

Commissioner & Secretary

Intermediate Education

ANDHRA PRADESH

GUNTUR

S.I.V.E Co – Ordinating Committee

Sri. Podili Yerraiah, **M.Sc., B.Ed.**
Professor
State Institute of Vocational Education
Commissioner of Intermediate Education, Guntur

Sri. B. Nageswara **B.Com, B.L.,**
Joint Secretary (Vocational)
Board of Intermediate Education, Vijayawada

Sri. Dr. G.V.S.R. Murthy, **M.Sc., Ph.D.**
Lecturer
State Institute of Vocational Education
Commissioner of Intermediate Education, Guntur

Text Book Development Committee

Physics

AUTHOR

K.S. Srinivasa Rao
M.Sc., B.Ed.,
Junior Lecturer in Physics
Government Junior College
Siripuram, Guntur

Chemistry

AUTHOR

G.J.V.S.N.D. Lakshmi
M.Sc.,
Junior Lecturer in Chemisty
SRR & CVR Jr. College
Vijayawada

EDITOR

Dr. R.V.S.S.N. Ravikumar
M.Sc., M.Phil., Ph.D., P.G.D.C.A.,
Professor & Chairman PG BOS in Physics
Acharya Nagarjuna University
Nagarjuna Nagar, Guntur-522510

Physics

Physics –II

CONTENTS

CHAPTER 1

WAVES

1.1 INTRODUCTION

We studied the motion of objects oscillating in isolation. What happens in a system, which is a collection of such objects? A material medium provides such an example. Here, elastic forces bind the constituents to each other and, therefore, the motion of one affects that of the other. If you drop a little pebble in a pond of still water, the water surface gets disturbed. The disturbance does not remain confined to one place, but propagates outward along a circle. If you continue dropping pebbles in the pond, you see circles rapidly moving outward from the point where the water surface is disturbed. It gives a feeling as if the water is moving outward iron the point of disturbance. If you put some cork pieces on the disturbed surface, it is seen that the cork pieces move up and down but do not move away from the centre of disturbance. This shows that the water mass does not flow outward with the circles, but rather a moving disturbance is created. Similarly, when we speak, the sound moves outward from us, without any flow of air from one part of the medium to another. The disturbances produced in air are much less obvious and only our ears or a microphone can select them. These patterns, which move without the actual physical transfer or flow of matter as a whole, are called waves. In this Chapter, we will study such waves.

Waves transport energy and the pattern of disturbance has information that propagates from one point to another. All our communications essentially depend on transmission of signals through waves. Speech means production of sound waves in air and hearing amounts to their detection. Often, communication involves different kinds of waves. For example, sound waves may be first converted into an electric current signal which in turn may generate an electromagnetic wave that may be transmitted by an optical cable or via a satellite. Detection of the original signal will usually involve these steps in reverse order.

Not all waves require a medium for their propagation. We know that light waves can travel through vacuum. The light emitted by stars, which are hundreds of light years away, reaches us through inter-stellar space, which is practically a vacuum.

The most familiar type of waves such as waves on a string, water waves, sound waves, seismic waves, etc., is the so-called mechanical waves. These waves require a medium for propagation, they cannot propagate through vacuum. They involve oscillations of constituent particles and depend on the elastic properties of the medium. The electromagnetic waves that you will learn in chapter XI are a different type of wave. Electromagnetic waves do not necessary require a medium – they can travel through vacuum. Light, radio waves, X-rays, are all electromagnetic waves. In vacuum, all electromagnetic waves have the same speed c, whose value is :

$$c = 299, 792, 458 \text{ ms}^{-1}. \quad (1.1)$$

A third kind of wave is the so-called Matter waves. They are associated with constituents of matter: electrons, protons, neutrons, atoms and molecules. They arise in quantum mechanical description of nature that you will learn in your later studies. Though conceptually more abstract than mechanical or electro-magnetic waves, they have already found applications in several devices basic to modern technology, matter waves associated with electrons are employed in electron microscopes.

In this chapter we will study mechanical waves, which require a material medium for their propagation.

The aesthetic influence of waves on art and literature is seen from very early times; yet the first scientific analysis of waves motion dates back to the seventeenth century. Some of the famous scientists associated with the physics of wave motion are Christian Huygens (1629 – 1695), Robert Hooke and Isaac Newton. The understanding of physics of waves followed the physics of oscillations of masses tied to springs and physics of the simple pendulum. Waves in elastic media are intimately connected with harmonic oscillations. (Stretched strings, coiled springs, air, etc., are examples of elastic media). We shall illustrate this connection through simple examples.

Consider a collection of springs connected to one another as shown in Fig 1.1. If the springs at one end is pulled suddenly and released, the disturbance travels to the other end. What has happened?

Fig. 1.1 A collection of springs connected to each other. The end A is pulled suddenly generating a disturbance, which then propagates to the other end.

The first spring is disturbed from its equilibrium length. Since the second spring is connected to the first, it is also stretched or compressed, and so on. The disturbance moves from one end to the other; but each springs only executes small oscillations about its equilibrium position. As a practical example of this situation, consider a stationary train at a railway station. Different bogies of the train are coupled to each other through a spring coupling. When an engine is attached at one end it gives a push to the bogie next to it; this push is transmitted from one bogie to another without the entire train being bodily displaced.

Now let us consider the propagation of sound waves in air. As the wave passes through air, it compresses or expands a small region of air. This causes a change in the density of that region, say, in that region. Pressure is force per unit area, so there is a **restoring force proportional** to the disturbance, just like in a spring. In this case, the quantity similar to extension or compression of the spring is the change in density. If a region is compressed, the molecules in that region are packed together, and they tend to move out to the adjoining region. Thereby increasing the density or creating compression in the adjoining. Consequently, the air in the first region undergoes rarefaction. If a region is comparatively rarefied the surrounding air will rush in making the rarefaction move to the adjoining region. Thus, the compression or rarefaction moves from one region to another, making the propagation of a disturbance possible in air.

In solids, similar arguments can be made. In a crystalline solid, atoms or group of atoms are arranged in a periodic lattice. In these, each atom or group of atoms is in equilibrium, due to forces from the surrounding atoms. Displacing one atom, keeping the others fixed, leads to restoring forces, exactly as in a spring. So we can think of atoms in a lattice as end points, with springs between pairs of them.

In the subsequent sections of this chapter we are going to discuss various characteristic properties of waves.

1.2 TRANVERSE AND LONGITUDINAL WAVES

We have seen that motion of mechanical waves involves oscillations of constituents of the medium. If the constituents of the medium oscillate perpendicular to the direction of wave propagation, we call the wave a transverse wave. If they oscillate along the direction of wave propagation, we call the wave a longitudinal wave.

Fig 1.2 shows the propagation of a single pulse along a string, resulting from a single up and down jerk. If the string is very long compared to the size of the pulse, the pulse will damp out before it reaches the other end and reflection from that end may be ignored. Fig 1.3 shows a similar situation, but this time the external agent gives a continuous periodic sinusoidal up and down jerk to one end of the string is then a sinusoidal wave. In either case the elements of the spring oscillate about their equilibrium mean position as the pulse or wave passes through them. The oscillations are normal to the direction of wave motion along the string, so this is an example of transverse wave.

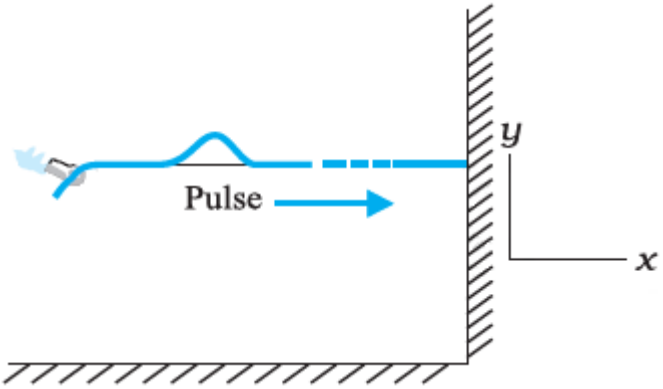


Fig. 1.2 When a pulse travels along the length of a stretched string (x-direction), the elements of the string oscillate up and down (y-direction)

We can look at a wave in two ways. We can fix an instant of time and picture the wave in space. This will give us the shape of the wave as a whole in space at a given instant. Another way is to fix a location i.e. fix our attention on a particular element of string and see its oscillatory motion in time.

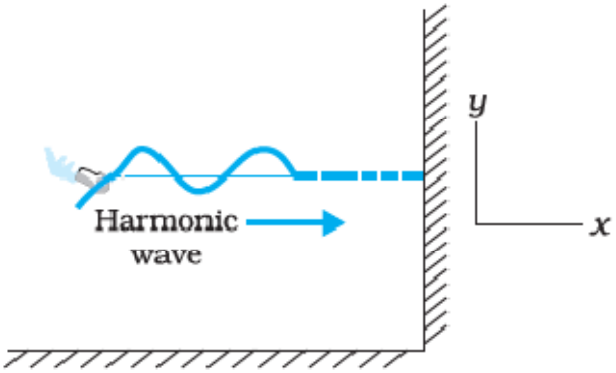


Fig. 1.3 A harmonic (sinusoidal) wave travelling along a stretched string is an example of a transverse wave. An element of the string in the region of the wave

oscillates about its equilibrium position perpendicular to the direction of wave propagation.

Fig. 1.4 Longitudinal waves (sound) generated in a pipe filled with air by moving the piston up and down. A volume element of air oscillates in the direction parallel to the direction of wave propagation.

Fig 1.4 describes the situation for longitudinal waves in the most familiar example of the propagation of sound waves. A long pipe filled with air has a piston at one end. A single sudden push forward and pull back of the piston will generate a pulse of condensation (higher density) and rarefactions (lower density) in the medium (air). If the push-pull of the piston is continuous and periodic (sinusoidal) a sinusoidal wave will be propagating in air along the length of the pipe. This is clearly an example of longitudinal waves.

The waves considered above, transverse or longitudinal, are travelling or progressive waves since they travel from one part of the medium to another. The material medium as a whole does not move, as already noted. A stream, for example, constitutes motion of water as a whole. In a water wave, it is the disturbance that moves, not water as a whole. Likewise a wind (motion of air as a whole) should not be confused with a sound wave which is a propagation of disturbance (in pressure density) in air, without the motion of air medium as a whole. Mechanical waves are related to the elastic properties of the medium. In transverse waves, the constituents of the medium oscillate perpendicular to wave motion causing change is shape. That is, each element of the medium in subject to shearing stresses. Solids and strings have shear modulus, which are they sustaining shearing stress. Fluids have no shape of their own – they yield to shearing stress. This is why transverse waves are possible in solids and strings (under tension) but not in fluids. However, solids as well fluids have bulk modulus, which is, they can sustain compressive strain. Since longitudinal waves involve compressive stress (pressure), they can be propagated through solids and fluids. Thus a steel bar possessing both bulk and sheer elastic moduli can propagate longitudinal as well as transverse waves. But air can propagate only longitudinal and transverse waves, their speeds can be different since they arise from different elastic moduli.

1.3 DISPLACEMENT RELATION IN A PROGRESSIVE WAVE

For mathematical description of a travelling wave, we need a function of both position x and time t. Such a function at every instant should give the shape of the wave at that instant. Also at every given location, it should describe the motion of the constituent of

the medium at that location. If we wish to describe a sinusoidal travelling wave, (such as the one shown in (Fig .1.3) the corresponding function must also be sinusoidal. For convenience, we shall take the wave to be transverse, we shall take the wave to be transverse so that if the position of the constituents of the medium is denoted by *x*, the displacement from the equilibrium position may be denoted by *y*. A sinusoidal travelling wave is then described by:

$$y(x,t) = a\sin\ (kx - \omega t + \emptyset) \qquad (1.2)$$

The term $\emptyset$ in the argument of sine function means equivalently that we are considering a linear combination of sine and cosine functions:

$$y(x,t) A\sin(kx - \omega t)\ B\ cos\ (kx - \omega t) \qquad (1.3)$$

From Equations (1.2) and (1.3)

$$a = \sqrt{A^2 + B^2}\ and\ tan^{-1}\left(\frac{B}{A}\right)$$

To understand why Equation (1.2) represents a sinusoidal travelling wave, take a fixed instant, say $t = t_0$. Then the argument of the sine function in Equation (1.2) is simply kx + constant. Thus the shape of the wave (at any fixed instant) as a function of x is a sine wave. Similarly, take a fixed location, say $x = x_0$. Then the argument of the sine function in Equation (1.2) is constant$-\omega t$. The displacement y, at a fixed location, thus varies sinusoidally with time. That is, the constituents of the medium at different positions execute simple harmonic motion. Finally, as t increases, x must increase in the positive direction to keep $kx - \omega t + \emptyset$ constant. Thus Eq.(1.2) represents a sinusoidal (harmonic) wave travelling along the positive direction of the x-axis. On the other hand, a function

$$\boldsymbol{y(x,t)\ a\sin(\ kx - \omega t + \emptyset)} \qquad (1.4)$$

Represent a wave travelling in the negative direction of x-axis. Fig (1.5) gives the names of the various physical quantities appearing in Eq. 1.2 that we now interpret.

$y(x,t)$	:	displacement as a function of position x and time t
a	:	amplitude of a wave
ω	:	angular frequency of the wave
k	:	angular wave number
$kx-\omega t+\phi$	:	initial phase angle ($a+x = 0$, $t = 0$)

Fig. 1.5 The meaning of standard symbols in Eq. (15.2)

Fig. 1.6 shows the plots of Eq. (1.2) for different values of time differing by equal interval of time. In a wave, the crest is the point of maximum positive displacement; the trough is the point of maximum negative displacement. To see how a wave travels, we can fix attention on a crest and see how it progresses with time. In the figure, this is shown by a cross (X) on the crest. In the same manner, we can see the motion of a particular constituent of the medium at a fixed location, say at the origin of the x-axis. This is shown by a solid dot (·).The plots of Fig.(1.6) show that with time, the solid(·)at the origin moves periodically i.e., the particle at the origin oscillates about its mean position as the wave progresses. This is true for any other location also.

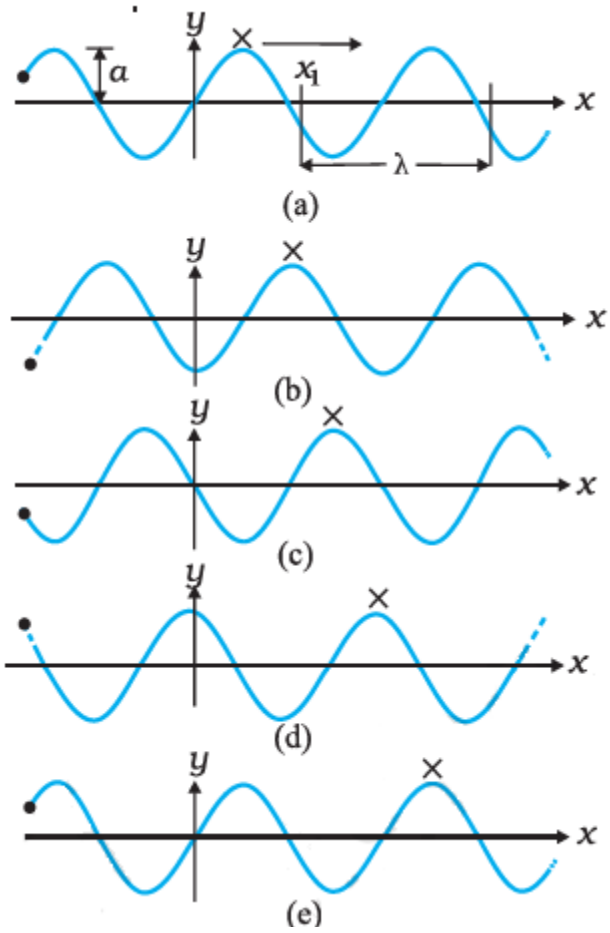


Fig. 1.6 A harmonic wave progressing along the positive direction of x-axis at different times.

We also see that during the time the solid dot (·) has completed one full oscillation, he crest has moved further by a certain distance. Using the plots of Fig (1.6) we now define the various quantities of Eq (1.2).

1.3.1 Amplitude and phase

In Eq.(1.2) since the sine function varies between 1 and -1, the displacement *y(x,t)* varies between ***a*** *and* ***–a***. We can take **a** to be a positive constant, without any loss of generality. Then ***a*** represents the maximum displacement of the constituents of the medium from their equilibrium position. Note that the displacement *y* may be positive or negative, but ***a*** is positive. It is called the **amplitude** of the wave.

The quantity $(kx - \omega t + \emptyset)$ appearing as the argument of the sine function in Eq (1.20 is called the phase of the wave. Given the amplitude *a*, the phase determines the displacement of the wave at any position and at any instant. Clearly $\emptyset$ is the phase at x = 0 and t = 0. Hence $\emptyset$ is called the initial phase angle. By suitable choice of origin on the x-axis and the initial time, it is possible to have $\emptyset = 0$. Thus there is no loss of generally in dropping $\emptyset$, i.e., in taking Eq (1.2) with $\emptyset = 0$.

1.3.2 Wavelength and Angular Wavenumber

The minimum distance between two points having the same phase is called the wave length of the wave, usually denoted by λ . For simplicity, we can choose points of the same phase to be crests or troughs. The wavelength is then the distance between two consecutive crests or troughs in a wave. Taking $\emptyset = 0$ in Eq.(1.2) the displacement at t = 0 is given by

$$\boldsymbol{y(x,0) = a \sin kx} \qquad (1.5)$$

Since the same function repeats its value after every 2π change in angle.

$$\sin kx \;\; \sin(kx + 2n\pi) = \sin k\left(x + \frac{2n\pi}{k}\right)$$

That is the displacements at points x and at $\boldsymbol{x + \frac{2n\pi}{k}}$ are the same, where n = 1, 2, 3 ..... . The least distance between points with the same displacement (at any given instant of time) is obtained by taking n = 1. λ is then given by

$$\boldsymbol{\lambda = \frac{2\pi}{k} \;\; or \; k = \frac{2\pi}{\lambda}} \qquad (1.6)$$

where k is the angular wave number or propagation constant; its SI unit is radian per metre or *Rad* m^{-1}

1.3.3 Period, Angular Frequency and Frequency

Fig 1.7 shows again a sinusoidal plot. It describes not the shape of the wave at a certain instant but the displacement of an element (at any fixed location) of the medium as a function of time. We may for, simplicity, take Eq.(1.2) with $\emptyset$ = 0 and monitor the motion of the element say at x = 0 . We then get

$$y\,(0,t) = \; a\sin(-\omega t) = -a\sin\omega t$$

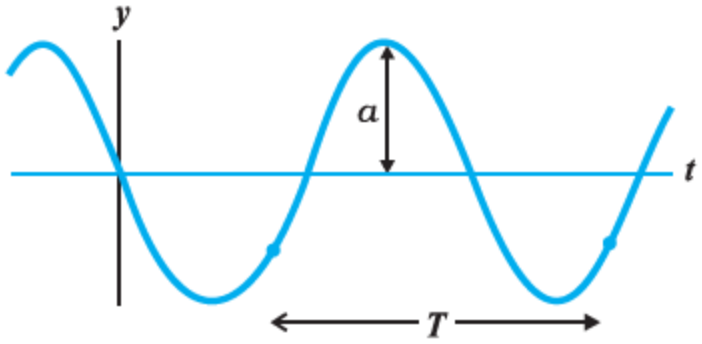


Fig. 1.7 An element of a string at a fixed location oscillates in time with amplitude a and period T, as the wave passes over it.

Now the period of oscillation of the wave is the time it takes for an element to complete one full oscillation.

That is

$$-a\sin\omega t = \; -a\sin\omega(t + T)$$
$$= \; -a\sin(\omega t + \; \omega T)$$

Since sine function repeats after every 2π .

$$\omega T = 2\pi \;\; or \;\; \omega = \frac{2\pi}{T} \qquad (1.7)$$

$\boldsymbol{\omega}$ is called the angular frequency of the wave. Its SI units is rad s^{-1} . The frequency v is the number of oscillations per second. Therefore

$$v = \frac{1}{T} = \frac{\omega}{2\pi} \qquad (1.8)$$

$\boldsymbol{v}$ **is** usually measured in hertz.

In the discussion above, reference has always been made to a wave travelling along a string or a transverse wave. In a longitudinal wave, the displacement of an element of the medium is parallel to the direction of propagation of the wave. In Eq (1.2) , the displacement function for a longitudinal wave is written as,

$$s\,(x,t) = a\sin(kx - \omega t + \phi) \qquad (1.9)$$

where s (x, t) is the displacement of an element of the medium in the direction of propagation of the wave at position x and time t. In Eq (1.9), a is the displacement amplitude: other quantities have the same meaning as in case of a transverse wave except that the displacement function $\boldsymbol{y\,(x,t)}$ is to be replaced by the function $\boldsymbol{s\,(x,t)}$.

1.4 THE SPEED OF A TRAVELLING WAVE

To determine the speed of propagation of a travelling wave, we can fix our attention on any particular point on the wave (characterized by some value of the phase) and see how that point moves in time. It is convenient to look at the motion of the crest of the wave. Fig (1.8) gives the shape of the wave at two instants of time which differ by a small time internal Δt. The entire wave pattern is seen to shift to the right (position direction of x-axis) by a distance Δx. In particular the crest shown by a cross (x) moves a distance Δx in time Δt.

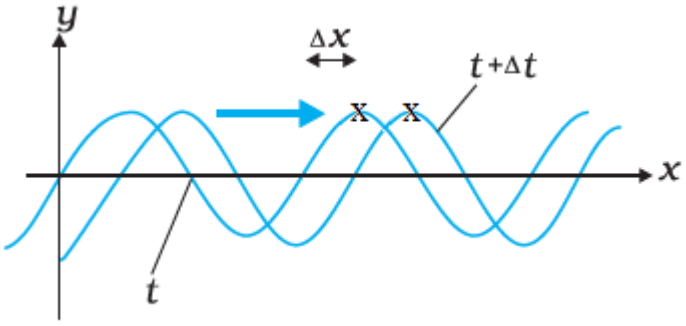


Fig. 1.8 Progression of a harmonic wave from time t to t + Δt. where Δt is a small interval. The wave pattern as a whole shifts to the right. The crest of the wave (or a point with any fixed phase) moves right by the distance Δx in time Δt.

The speed of the wave is then $\Delta x/\Delta t$. We can put the cross (x) on a point with any other phase. It will move with the same speed v (other wise the wave pattern will not remain fixed). The motion of a fixed phase point on the wave is given by

$$kx - \omega t = constant \qquad (1.10)$$

Thus, as time t changes, the position x of the fixed phase point must change so that the phase remains constant . Thus

$$kx - \omega t = k\,(x + \Delta x) - \omega\,(t + \Delta t)$$

$$\text{Or } k\ \Delta x - \omega \Delta t = 0$$

Taking Δx , Δt vanishingly small, this gives

$$\frac{dx}{dt} = \frac{\omega}{k} = v \quad (1.11)$$

Relating ω to T and k to $\boldsymbol{\lambda}$, we get

$$v = \frac{2\pi\omega}{2\pi k} = \lambda v = \frac{\lambda}{T} \quad (1.12)$$

Eq. (1.12), a general relation for all progressive waves, shows that in the time required for one full oscillation by any constituent of the medium, the wave pattern travels a distance equal to the wavelength of the wave. It should be noted that the speed of a mechanical wave is determined by the inertial (linear mass density for strings , mass density in general) and elastic properties (Young's modulus) of the medium. The medium determines the speed ; Eq (1.12) then relates wavelength to frequency for the given speed. Ofcourse, as remarked earlier, the medium can support both transverse and longitudinal waves, which will have different speeds in the same medium. Later in this chapter, we shall obtain specific expressions for the speed of mechanical waves in some media.

1.4.1 Speed of Transverse Wave on Stretched String

The speed of a mechanical wave is determined by the restoring force setup in the medium when it is disturbed and the inertial properties (mass density) of the medium. The speed is expected to be directly related to the former and inversely to the latter. For waves on a string, the restoring force is provided by the tension T in the string. The inertial property will in this case be linear mass density μ, which is mass m of the string divided by its length L. Using Newton's Laws of Motion, an exact formula for the wave speed on a string can be derived, but this derivation is outside the scope of this book. We shall, therefore use dimensional analysis. We already know that dimensional analysis alone can never yield the exact formula. The overall dimensionless constant is always left undetermined by dimensional analysis.

The dimension of μ is $[ML^{-1}]$ and that of T is like force, namely $[MLT^{-2}]$. We need to combine these dimensions to get the dimension of speed v $[LT^{-1}]$. Simply inspection shows that the quantity T/μ has the relevant dimension.

$$\frac{MLT^2}{ML^{-1}} = L^2T^2$$

Thus if T and μ are assumed to be the only relevant physical quantities.

$$v = C\sqrt{\frac{T}{\mu}} \quad (1.13)$$

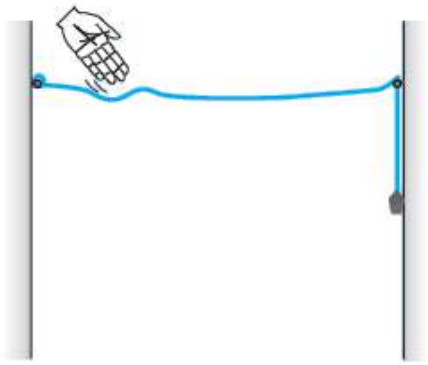
Where C is the undetermined constant of dimensional analysis. In the exact formula. It turns out, C = 1. The speed of transverse waves on a stretched string is given by

$$v = \sqrt{\frac{T}{\mu}} \quad (1.14)$$

Note the important point that the speed v depends only on the properties of the medium T and μ (T is a property of the stretched string arising due to an external force). It does not depend on wave length or frequency of the wave itself. In higher studies, you will come across waves whose speed is not independent of frequency of the wave. Of the two parameters λ and v the source of disturbance determines the frequency of the wave generated. Given the speed of the wave in the medium and the frequency Eq. (1.12) then fixes the wavelength

$$\lambda = \frac{v}{\nu} \quad (1.15)$$

Propagation of a pulse on a rope



You can easily see the motion of a pulse on a rope. You can also see its reflection from a rigid boundary and measure its velocity of travel. You will need a rope of diameter 1 to 3 cm, two hooks and some weights. You can perform this experiment in your classroom or laboratory.

Take a long rope or thick string of diameter 1 to 3 cm, and tie it to hooks on opposite walls in a hall or laboratory. Let one end pass on a hook and hang some weight (about 1 to 5 kg) to it. The walls may be about 3 to 5 m apart.

Take a stick or a rod and strike the rope hard at a point near one end. This creates a pulse on the rope which now travels on it. You can see it reaching the end and reflecting back from it. You can check the phase relation between the incident pulse and reflected pulse. You can easily watch two or three reflections before the pulse dies out. You can take a stopwatch and find the time for the pulse to travel the distance between the walls, and thus measure its velocity. Compare it with that obtained from Eq. (1.14). This is also what happens with a thin metallic string of a musical instrument. The major difference is that the velocity on a string is fairly high because of low mass per unit length, as compared to that on a thick rope. The low velocity on a rope allows us to watch the motion and make measurements beautifully.

1.4.2 Speed of a Longitudinal Wave (Speed of Sound)

In a longitudinal wave the constituents of the medium oscillate forward and backward in the direction of propagation of the wave. We have already seen that the sound waves travel in the form of compressions and rarefactions of small volume elements of air. The elastic property that determines the stress under compressional strain is the bulk modulus of the medium defined by

$$B = -\frac{\Delta P}{\Delta V / V} \tag{1.16}$$

Here the change in pressure ΔP produces a volumetric strain $\frac{\Delta V}{V}$, B has the same dimension as pressure and given in SI units in terms of pascal (Pa) . The inertial property relevant for the propagation of wave in the mass density , with dimensions [ML^{-3}]. Simple inspection reveals that quantity B/ ρ has the relevant dimension :

$$\frac{[ML^{-1}T^{-2}]}{[ML^{-3}]} = [L^2T^{-2}] \tag{1.17}$$

Thus if B and ρ are considered to be the only relevant physical quantities.

$$v = C\sqrt{\frac{B}{\rho}} \tag{1.18}$$

Where, as before, C is the undetermined constant from dimensional analysis. The exact derivation shows that C = 1. Thus the general formula for longitudinal waves in a medium is:

$$v = \sqrt{\frac{B}{\rho}} \tag{1.19}$$

For a linear medium like a solid bar, the lateral expansion of the bar is negligible and we may consider it to be only under longitudinal strain. In that case, the relevant modulus of elasticity in Young's modulus, which has the same dimension as the Bulk modulus. Dimensional analysis for this case is the same as before and yields a relation like Eq (1.18), with an undetermined C which the exact derivation shows to be unity. Thus the speed of longitudinal waves in a solid bar is given by

$$v = \sqrt{\frac{Y}{\rho}} \qquad (1.20)$$

where Y is the Young's modulus of the material of the bar. Table 1.1 gives the speed of sound in some media.

Table 1.1 Speed of Sound in some Media

Medium	**Speed (m s^{-1})**
Gases	
Air (0 ^oC)	331
Air (20 ^oC)	343
Helium	965
Hydrogen	1284
Liquids	
Water (0 ^oC)	1402
Water (20 ^oC)	1482
Seawater	1522
Solids	
Aluminium	6420
Copper	3560
Steel	5541
Granite	6000
Vulcanised Rubber	54

Liquids and solids generally have higher speeds of sound than in gases. [Note for solids, the speed being referred to is the speed of longitudinal waves in the solid]. This happens because they are much more difficult to compress than gases and so have much higher values of bulk modulus. This factor more than compensates for their higher densities than gases.

We can estimate the speed of sound in a gas in the ideal gas approximation. For an ideal gas, the pressure P, volume V and temperature T are related by (see Chapter 12 of 1st Yr. book).

$$PV = NK_BT \qquad (1.21)$$

Where N is the number of molecules in volume V. K_B is the Boltzmann constant and T the temperature of the gas (in Kelvin) . Therefore , for an isothermal change it follows from Eq.(1.21) that

$$V\Delta P + P\Delta V = 0$$

Or $$-\frac{\Delta P}{\Delta V/V} = P$$

Hence, substituting in Eq. (1.16), we have

$$B = P$$

Therefore, from Eq (1.19) the speed of a longitudinal wave in an ideal gas is given by ,

$$v = \sqrt{\frac{P}{\rho}} \qquad (1.22)$$

This relation was first given by Newton and is known as Newton's formula.

If we examine the basic assumption made by Newton that the pressure variations in a medium during propagation of sound are isothermal, we find that this is not correct. It was pointed out by Laplace that the pressure variations in the propagation of sound waves are so fast that there is little time for the heat flow to maintain constant temperature. These variations, therefore, are adiabatic and not isothermal. For adiabatic processes the ideal gas satisfies the relation.

$$PV^{\gamma} = constant$$

i.e., $$\Delta(PV^{\gamma}) = 0$$

or $$P\gamma V^{\gamma-1}\Delta V + V^{\gamma}\Delta P = 0$$

Thus, for an ideal gas the adiabatic bulk modulus is given by,

$$B_{ad} = -\frac{\Delta P}{\Delta V/V} = \gamma P$$

Where γ is the ratio of two specific heats, C_P/C_v. The speed of sound is, therefore, given by,

$$\boldsymbol{v = \sqrt{\frac{\gamma P}{\rho}}} \qquad \textbf{(1.23)}$$

This modification of Newton's formula is referred to as the **Laplace correction**. For air $\gamma = 7/5$.

Now using Eq.(1.23) to estimate the speed of sound in air at STP, we get a value 331.3 m s^{-1} which agrees with the measured speed.

1.5 THE PRINCIPLE OF SUPERPOSITION OF WAVES

What happens when two wave pulses travelling in opposite directions cross each other ? It turns out that wave pulses continue to retain their identities after they have crossed. However , during the time they overlap, the wave pattern is different from either of the overlap. Figure 1.9 shows the situation when two pulses of equal and opposite shapes move towards each other. When the pulses overlap, the resultant displacement due to each pulse. This is known as the principle of superposition of waves. According to this principle, each pulse moves as if others are not present. The constituents of the medium therefore suffer displacements due to both and since displacements can be positive and negative, the net displacement is an algebraic sum of the two. Fig (1.9) gives graphs of the wave shape at different times. Note the dramatic effect in the graph (c); the displacements due to the two pulses have exactly cancelled each other and there is zero displacement throughout.

To put the principle of superposition mathematically, let $y_1(x,t) and\ y_2(x,t)$ be the displacements due to two waves disturbances in the medium.

If the waves arrive in a region simultaneously and therefore, overlap, the net displacement, *y(x, t)* is given by

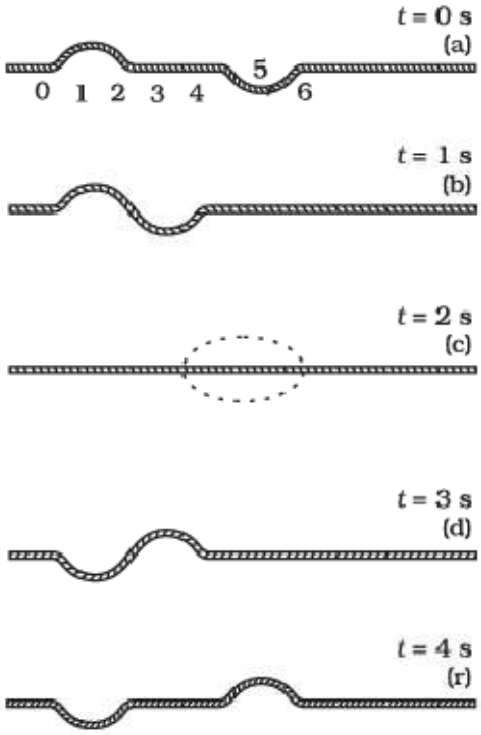


Fig. 1.9 Two pulses having equal and opposite displacements moving in opposite directions. The overlapping pulses add up to zero displacement in curve (c).

$$y\,(x, t) = y_1(x, t) + y_2(x, t) \qquad (1.24)$$

If we have two or more waves moving in the medium the resultant waveform is the sum of wave functions of individual waves. That is, if the wave functions of the moving waves are

$$y_1 = f_1(x - vt)\,,$$

$$y_2 = f_2(x - vt)\,,$$

.....

.....

$$y_n = f_n(x - vt)$$

Then the wave function describing the disturbance in the medium is

$$y = f_1(x - vt) + f_2(x - vt) + \cdots\ldots\ldots + f_n(x - vt)$$
$$= \sum_{i=1}^{n} f_i(x - vt) \qquad (1.25)$$

The principle of superposition is basic to the phenomenon of interference.

For simplicity, consider two harmonic travelling waves on a stretched string, both with the same ω (angular frequens) and k (wave number), and, therefore, the same wavelength λ. Their wave speeds will be identical .Let us further assume that their amplitudes are equal and they are both travelling in the positive direction of x-axis. The

waves only differ in their initial phase. According to Eq.(1.2) , the two waves are described by the functions :

$$y_1(x,t) = a\sin(kx - \omega t) \quad (1.26)$$

and $$y_2(x,t) = a\,sin\,(\,kx - \omega t + \phi) \quad (1.27)$$

The net displacement is then, by the principle of superposition, given by

$$y(x,t) = a\sin(kx - \omega t) + a\sin(\,kx - \omega t + \phi) \quad (1.28)$$

$$= a\left[2\sin\left[\frac{(kx-\omega t)+(kx-\omega t+\phi}{2}\right]cos\frac{\phi}{2}\right] \quad (1.29)$$

Where we have used the familiar trigonometric identity for sin A + sin B. We then have

$$y\,(x,t) = 2a\,cos\frac{\phi}{2}\sin\left(kx - \omega t + \frac{\phi}{2}\right) \quad (1.30)$$

Eq. (1.30) is also a harmonic travelling wave in the positive direction of x-axis, with the same frequency and wave length. However, its initial phase angle is $\frac{\phi}{2}$**.** The significant thing is that its amplitude is a function of the phase difference ϕ between the constituent two waves:

$$A(\phi) = 2a\cos(\phi/2) \quad (1.31)$$

For ϕ = 0, when the waves are in phase.

$$y\,(x,t) = 2a\sin(kx - \omega t) \quad (1.32)$$

i.e., the resultant wave has amplitude 2a, the largest possible value for A. For $\phi = \pi$, the waves are completely, out of phase and the resultant wave has zero displacement everywhere at all times

$$y\,(\,x,t) = 0 \quad (1.33)$$

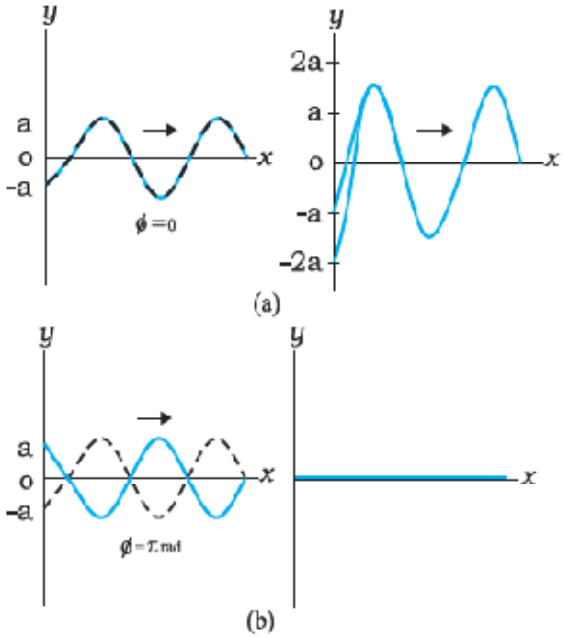


Fig. 1.10 The resultant of two harmonic waves of equal amplitude and wavelength according to the principle of superposition. The amplitude of the resultant wave depends on the phase difference ϕ, which is zero for (a) and π for (b).

Eq. (1.32) refers to the so-called constructive interference of the two waves where the amplitudes add up in the resultant wave. Eq. (1.33) is the case of destructive interference where the amplitudes subtract out in the resultant wave. Fig (1.10) shows these two cases of interference of waves arising from the principle of superposition.

1.6 REFLECTION OF WAVES

So far we considered waves propagating in an unbounded medium. What happens if a pulse or a wave meets a boundary? If the boundary is rigid, the pulse or wave gets reflected. The phenomenon of echo is an example of reflection by a rigid boundary. If the boundary is not completely rigid or is an interface between two different elastic media, the situation is some what complicated. A part of the incident wave is reflected and a part is transmitted into the second medium.

If a wave is incident obliquely on the boundary between two different media the transmitted wave is called the **refracted wave.** The incident and refracted waves obey Snell's law of refraction, and the incident and reflected waves obey the usual laws of reflection.

Fig (1.11) shows a pulse travelling along a stretched string and being reflected by the boundary. Assuming there is no absorption of energy by the boundary, the reflected wave has the same shape as the incident pulse but it suffers a phase change of π or 180^0 on reflection. This is because the boundary is rigid and the disturbance must have zero displacement at all times at the boundary. By the principle of superposition, this is possible only if the reflected and incident waves differ by a phase of π, so that the resultant displacement is zero. This reasoning is based on boundary condition on a rigid wall. We can arrive at the same conclusion dynamically also. As the pulse arrives at the wall, it exerts a force on the wall. By Newton's Third Law, the wall exerts an equal and opposite force on the string generating a reflected pulse that differs by a phase of π .

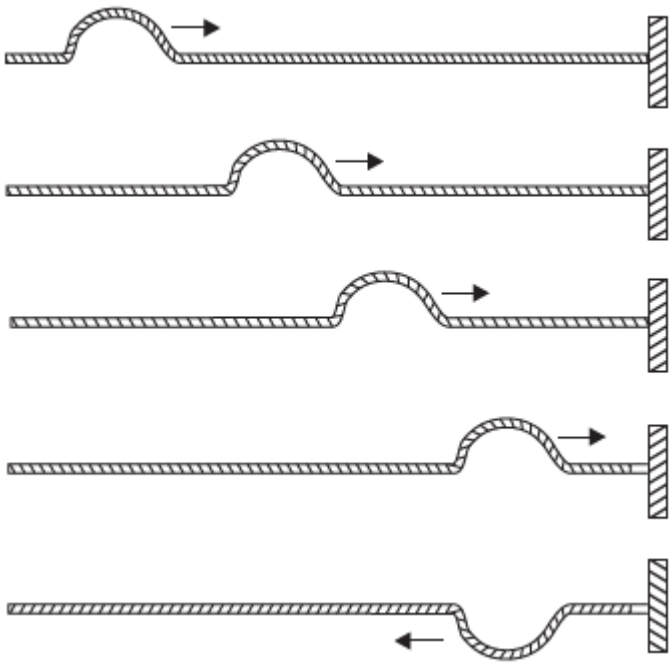


Fig. 1.11 Reflection of a pulse meeting a rigid boundary.

If on the other hand, the boundary point is not rigid but completely free to move (such as in the case of a string tied to a freely moving ring on a rod), the reflected pulse has the same phase and amplitude (assuming no energy dissipation) as the incident pulse. The net maximum displacement at the boundary is then twice the amplitude of each pulse. An example of non – rigid boundary is the open end of an organ pipe.

To summarize, a travelling wave or pulse suffers a phase change of π on reflection at a rigid boundary at an open boundary. To put this mathematically, let the incident travelling wave be

$$y_2(x,t) = a\sin(kx - \omega t)$$

At a rigid boundary, the reflected wave is given by

$$y_1(x,t) = a\sin(kx - \omega t + \pi) .$$
$$= -a\ sin\ (kx - \omega t) \quad (1.34)$$

At an open boundary, the reflected wave is given by

$$y_1(x,t) = a\sin(kx - \omega t + 0).$$
$$= a\ sin\ (kx - \omega t) \quad (1.35)$$

Clearly, at he rigid boundary, $y = y_2 + y_r = 0$ at all times.

1.6.1 Standing Waves and Normal Modes

We considered above reflection at one boundary. But there are familiar situations (a string fixed at either and or an ait column in a pipe with either and closed) in which reflection takes place at two or more boundaries . In a string for example, a wave going to the right will get reflected at one end, which in turn will travel and get reflected from the other end. This will go on until there is a steady wave pattern set up on the string. Such wave patterns are called standing waves or stationary waves. To see this mathematically, consider a wave travelling along the positive direction of x-axis and a reflected wave of the same amplitude and wavelength in the negative direction of x-axis. From Eqs. (1.2) and (1.4) with $\emptyset = 0$, we get :

$$y_1(x,t) = a\sin(kx - \omega t)$$
$$y_2(x,t) = a\sin(kx + \omega t)$$

The resultant wave on the string is, according to the principle of superposition:

$$y(x,t) = y_1(x,t) + y_2(x,t)$$
$$= a[\sin(kx - \omega t) + a\sin(kx + \omega t)]$$

Using the familiar trigonometric identity Sin (A + B) + Sin (A – B) = 2 sin A cos B we get,

$$y(x,t) = 2a\ \sin kx \cos \omega t \quad (1.36)$$

Note the important difference in the wave pattern described by Eq. (1.36) from that described by Eq.(1.2) or Eq.(1.4). The terms kx and ωt appear separately, not in the combination kx – ωt . The amplitude of this wave is *2a sin kx* . Thus in this wave pattern, the amplitude varies from point to point, but each element of the spring oscillates with the same angular frequency ω or time period. There is no phase difference between oscillations of different elements of the wave. The spring as a whole vibrates in phase with differing amplitudes at different points. The wave pattern is neither moving to the right nor to the left . Hence they are called standing or stationary waves. The amplitude is fixed at a given location

but, as remarked earlier, it is different at different locations. The points at which the amplitude is zero (i.e., where there is no motion at all) are **nodes;** the points at which the amplitude is the largest are called **antinodes.** Fig (1.12) shows a stationary wave pattern resulting from superposition of two travelling waves in opposite directions.

The most significant feature of stationary waves is that the boundary conditions constrain the possible wavelengths or frequencies of vibration of the system. The system cannot oscillate with any arbitrary frequency (contrast this with a harmonic travelling wave), but is characterized by a set of natural frequencies or **normal modes** of oscillation. Let us determine these normal modes for a stretched string fixed at both ends.

First , from Eq.(1.36) , the positions of modes (where the amplitude is zero) are given by *sin kx = 0*. which implies

$$kx = n\pi \, , n = 0, 1, 2, 3, \ldots \ldots \ldots$$

Since $= 2\pi/\lambda$, we get

$$x = \frac{n\lambda}{2} \; ; \; n = 0, 1, 2, 3, \ldots \ldots \ldots \qquad (1.37)$$

Clearly , the distance between any two successive nodes is $\frac{\lambda}{2}$. In the same way, the positions of antinodes (where the amplitude is the largest) are given by the largest value of *sin kx* .

$$|\sin kx| = 1$$

Which implies $kx = \left(n + \frac{1}{2}\right)\pi \; ; n = 0, 1, 2, 3, \ldots \ldots \ldots \ldots$

With $= 2\pi/\lambda$, we get

$$x = \left(n + \frac{1}{2}\right)\frac{\lambda}{2} \; ; n = 0, 1, 2, 3, \ldots \ldots \ldots \qquad (1.38)$$

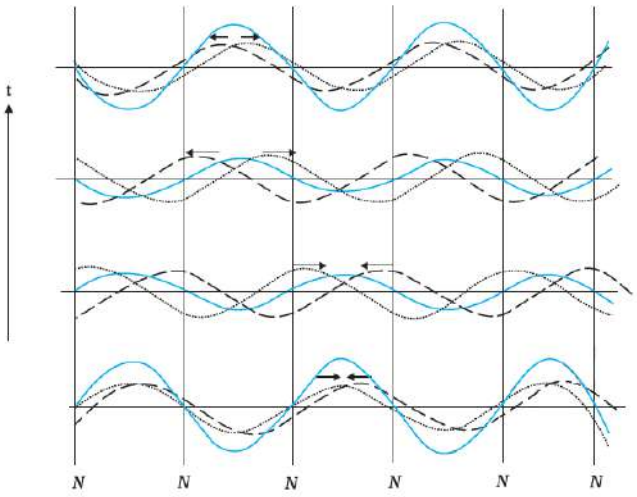


Fig. 1.12 Stationary waves arising from superposition of two harmonic waves travelling in opposite directions. Note that the positions of zero displacement (nodes) remain fixed at all times.

Again the distance between any two consecutive antinodes is $\frac{\lambda}{2}$. Eq.(1.37) can be applied to the case of a stretched string of length L fixed sat both ends. Taking one end to be

at x = 0, the boundary conditions are that x = 0 and x = L are positions of nodes. The x = 0 condition is already satisfied.

The x = L node condition requires that the length L is related to $\boldsymbol{\lambda}$ by

$$L = n\ \frac{\lambda}{2}\ ; \quad n = 1, 2, 3, \ldots \ldots \ldots \tag{1.39}$$

Thus, the possible wavelengths of stationary waves are constrained by the relation

$$\lambda = \frac{2L}{n}\ ;\ n = 1, 2, 3, \ldots \ldots \ldots \ldots \tag{1.40}$$

With corresponding frequencies

$$v = \frac{nv}{2L}\ , for\ n = 1, 2, 3, \ldots \ldots \ldots \ldots \tag{1.41}$$

We have thus obtained the natural frequencies the normal modes of oscillation of the system. The lowest possible natural frequency of a system is called its **fundamental mode** or the **first harmonic.**

For the stretched string fixed at either and it is given by $\boldsymbol{v = \frac{v}{2L}}$**,** corresponding to n = 1 of Eq .(1.41). Here v is the speed of wave determined by the properties of the medium. The n = 2 frequency is called the second harmonic; n = 3 is the third harmonic and so on. We can label the various harmonics by the symbol v_n (n = 1, 2 ,....).

Fig (1.13) shows the first six harmonics of a stretched string fixed at either end. A string need not vibrate in one of these modes only. Generally, the vibration of a string will be a superposition of different modes; some modes may be more strongly excited and some less. Musical instruments like sitar or violin are based on this principal. Where the string is plucked or bowed, determines which modes are more prominent than others.

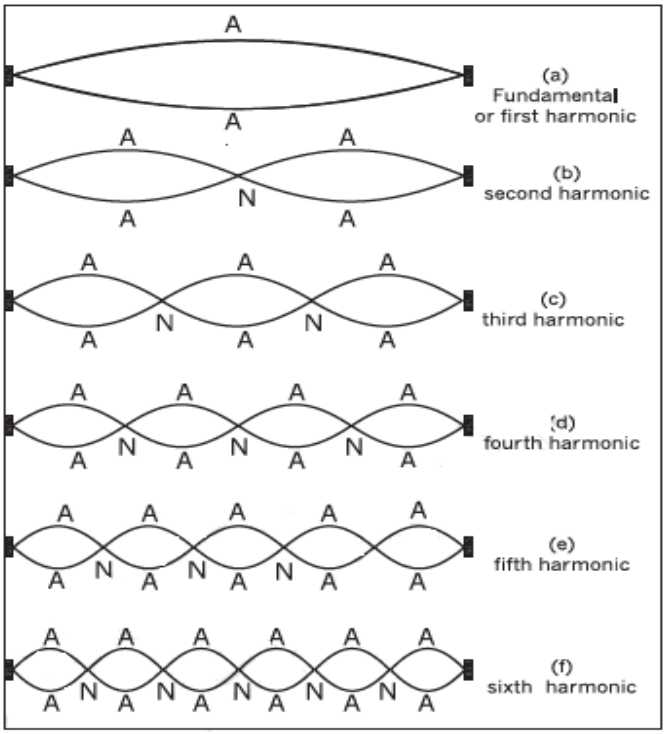


Fig. 1.13 The first six harmonics of vibrations of a stretched string fixed at both ends.

Let us next consider normal modes of oscillation of an air column with one end closed and the other open. A glass tube partially filled with water illustrates this system. The end in contact with water is a node, while the open end is an antinode . At the node the pressure changes are the largest, while the displacement is minimum (zero) . At the open end – the antinode, it is just the other way – least pressure change and maximum amplitude of displacement. Taking the end in contact with water to be x = 0 , the node condition (Eq.(1.37) is already satisfied . If the other end x = L is an antinode, Eq (1.38) gives

$$L = \left(n + \frac{1}{2}\right)\frac{\lambda}{2} \text{ , } for\ n = 0, 1, 2, 3, \ldots \ldots \ldots$$

The possible wavelengths are then restricted by the relation:

$$\lambda = \frac{2L}{(n+1/2)}, for\ n = 0, 1, 2, 3, \ldots \ldots \ldots \qquad (1.42)$$

The normal modes – the natural frequencies – of the system are

$$v = \left(n + \frac{1}{2}\right)\frac{v}{2L} \text{ ; } n = 0, 1, 2, 3, \ldots \ldots \ldots \qquad (1.43)$$

The fundamental frequency corresponds to n = 0, and is given by $\frac{v}{4L}$. The higher frequencies are odd harmonics, i.e., odd multiples of the fundamental frequency: $3\frac{v}{4L}$, $5\frac{v}{4L}$, etc.

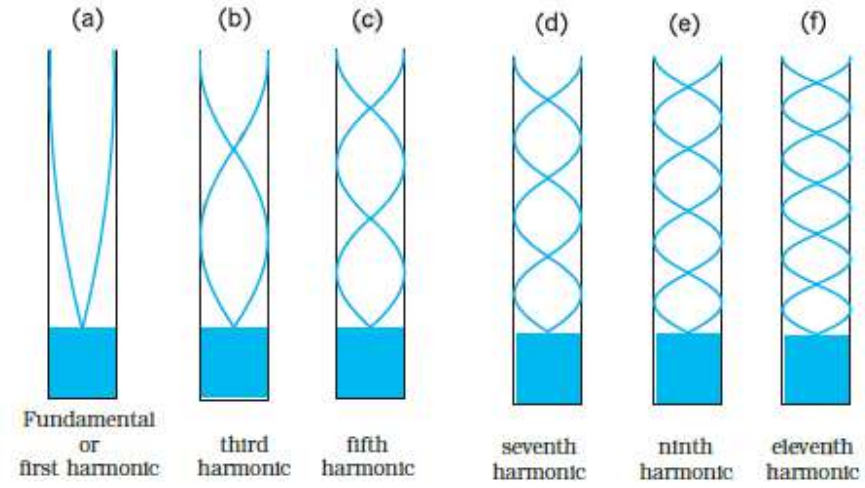


Fig. 1.14 Normal modes of an air column open at one end and closed at the other end. Onlythe odd harmonics are seen to be possible.

Fig. 1.14 shows the first six odd harmonics of air column with one end closed and the other open. For a pipe open at both ends, each end is an antinode. It is then easily seen that an open air column at both ends generates all harmonics [See Fig (1.15)].

The systems above, strings and sir columns, can also undergo forced oscillations (Chapter – 8 of I year) .If the external frequency is close to one of the natural frequencies, the system shows resonance**.**

Normal modes of a circular membrane rigidly clamped to the circumference as in a table are determined by the boundary condition that no point on the circumference of the membrane vibrates. Estimation of the frequencies of normal modes of this system is more complex. This problem involves wave propagation in two dimensions. However, the underlying physics is the same.

1.7 BEATS

'Beats' is an interesting phenomenon arising from interference of waves. When two harmonic sound waves of close (but not equal) frequencies are heard at the same time, we hear a sound of similar frequency (the average of two close frequencies), but we hear something else also.

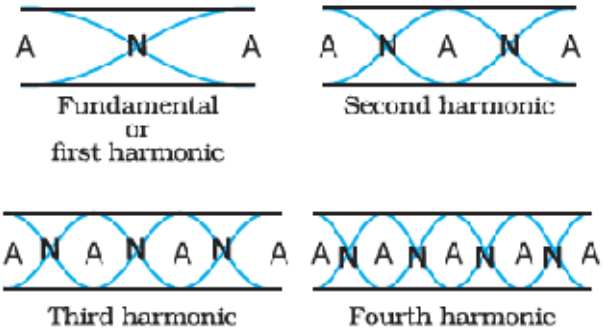


Fig. 1.15 Standing waves in an open pipe, first four harmonics are depicted.

We hear audibly distinct waxing and waning of the intensity of the sound, with a frequency equal to the difference in the two close frequencies. Artists use this phenomenon often while tuning their instruments with each other. They go on tuning until their sensitive ears do not detect any beats.

To see this mathematically, let us consider two harmonic sound waves of nearly equal angular frequency ω_1 and ω_2 and fix the location to be x = 0 for convenience. Eq.(1.2) with a suitable choice of phase ($\phi = \pi/2$ for each) and, assuming equal amplitudes, given

$$s_1 = a\cos\omega_1 t \quad and \quad s_2 = a\cos\omega_2 t \tag{1.44}$$

Here we have replaced the symbol y by s, since we are referring to longitudinal not transverse displacement. Let $\boldsymbol{\omega_1}$ be the (slightly) greater of the two frequencies.

Musical Pillars

Temples often have some pillars portraying human figures playing musical instruments, but seldom do these pillars themselves produce music. At the Nellaiappar temple in Tamil Nadu, gentle taps on a cluster of pillars carved out of a single piece of rock produce the basic notes of Indian classical music, viz. Sa, Re, Ga, Ma, Pa, Dha, Ni, Sa. Vibrations of these pillars depend on elasticity of the stone used, its density and shape.

Musical pillars are categorised into three types: The first is called the Shruti Pillar, as it can produce the basic notes — the "swaras". The second type is the Gana Thoongal, which generates the basic tunes that make up the "ragas". The third variety is the Laya Thoongal pillars that produce "taal" (beats) when tapped. The pillars at the Nellaiappar temple are a combination of the Shruti and Laya types. Archaeologists date the Nelliappar temple to the 7th century and claim it was built by successive rulers of the Pandyan dynasty.

The musical pillars of Nelliappar and several other temples in southern India like those at Hampi (picture), Kanyakumari, and Thiruvananthapuram are unique to the country and have no parallel in any other part of the world.

The resultant displacement is, by the principle of superposition.

$$s = s_1 + s_2 = a\,(\cos\omega_1 t + \cos\omega_2 t)$$

Using the familiar trigonometric identity for cos A + cos B , we get

$$= 2\,a\cos\frac{(\omega_1-\omega_2)t}{2}\cos\frac{(\omega_1+\omega_2)t}{2} \qquad (1.45)$$

Which may be written as :

$$S = [2\ a\cos\omega_b t]\ \cos\omega_a t \qquad (1.46)$$

If $|\omega_1 - \omega_2| \ll \omega_1, \omega_2, \omega_a \gg \omega_b$, th

Where $\omega_b = \frac{(\omega_1-\omega_2)}{2}$ and $\omega_a = \frac{(\omega_1+\omega_2)}{2}$

Now if we assume $|\omega_1 - \omega_2| \ll \omega_1$, which means $\omega_a \gg \omega_b$, we can interpret Eq. (1.46) as follows. The resultant wave is oscillating with the average angular frequency ω_a ; however its amplitude is not constant in time, unlike a pure harmonic wave. The amplitude is the largest when the term cos ω_b t takes its limit +1 or -1 . In other words, the intensity of the resultant wave waxes and wanes with a frequency which is 2 $\omega_b = \omega_1 - \omega_2$.

Since $\omega = 2\pi v$, the beat frequency v_{beat} , is given by

$$v_{beat} = v_1 - v_2 \qquad (1.47)$$

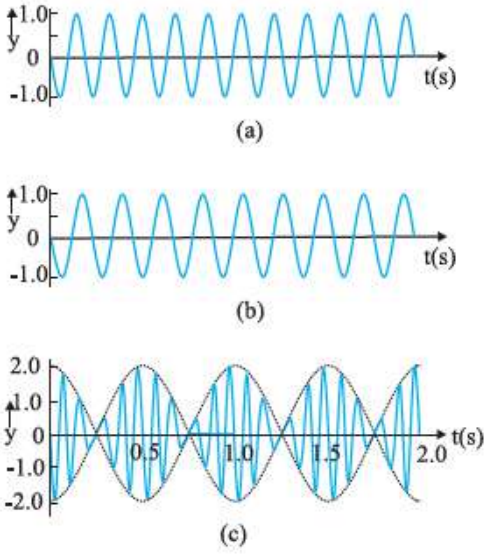


Fig. 1.16 Superposition of two harmonic waves, one of frequency 11 Hz (a), and the other of frequency 9Hz (b), giving rise to beats of frequency 2 Hz, as shown in (c).

Fig 1.16 illustrates the phenomenon of beats for two harmonic waves of frequencies 11 Hz and 9 Hz. The amplitude of the resultant wave shows beats at an frequency of 2 Hz.

1.8 DOPPLER EFFECT

It is an everyday experience that the pitch (or frequency) of the whistle of a fast moving train decreases as it recedes away. When we approach a stationary source of sound with high speed, the pitch of the sound heard appears to be higher than that of the source. As the observer recedes away from the source. As the observed **pitch** (or frequency) becomes lower than that of the source. This motion-related frequency change is called **Doppler effect**. The Austrian physicist Johann Christian Doppler first proposed the effect in 1842. Buys Ballot in Holland tested it experimentally in 1845. Doppler effect is a wave phenomenon, it holds not only for sound waves but also for electromagnetic waves. However, here we shall consider only sound waves.

We shall analyse changes in frequency under three different situations: (1) observer is stationary but the source is moving, (2) observer is moving but the source is stationary, and (3) both the observer and the source are moving . The situations (1) and (2) differ from each other because of the absence or presence of relative motion between the observer and the medium. Most waves require a medium for their propagation; however, electromagnetic waves do not require any medium for propagation. If there is no medium present, the Doppler shifts are same irrespective of whether the source moves or the observer moves, since there is no way of distinction between the two situations.

1.8.1 Source Moving: Observer Stationary

Let us choose the convention to take the direction from the observer to the source as the positive direction of velocity.

Consider a source S moving with velocity v_s and an observer who is stationary in a frame in which the medium is also at rest. Let the speed of a wave on angular frequency ω and period T_0, both measured by an observer at rest with respect to the medium, be v . We assume that the observer has a detector that counts every time a wave crest reaches it. As shown in Fig (1.17) at time t = 0 the source is at point S_1, located at a distance L from the observer, and emits a crest. This reaches the observer at time $t_1 = L/v$. At time t = T_0 the source has moved a distance $v_s T_0$ and is at point S_2, located at a distance (L + $v_s T_0$) from the observer.

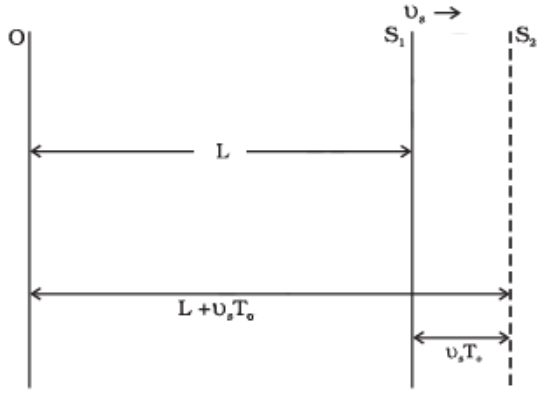


Fig. 1.17 Doppler effect (change in frequency of wave) detected when the source is moving and the observer is at rest in the medium.

At S_2 , the source emits a second crest. This reaches the observer at

$$t_2 = T_0 + \frac{(L+v_s T_0)}{v}$$

At time T_0 , the source emits its (n +1)th crest and this reaches the observer at time

$$t_{n+1} = n\,T_0 + \frac{(L+nv_sT_0)}{v}$$

Hence , in a time interval $\left[nT_0 = \frac{(L+nv_sT_0)}{v} - \frac{L}{v}\right]$ the observer's detector counts n crests and the observer records the period of the wave as T given by

$$T = \left[nT_0 + \frac{(L+nv_sT_0)}{v} - \frac{L}{v}\right] / \text{n}$$

$$= T_0 + \frac{v_sT_0}{v}$$

$$= T_0\left(1 + \frac{v_s}{v}\right) \qquad (1.48)$$

Equation (1.48) may be rewritten in terms of the frequency v_0 that would be measured if the source and observer were stationary , and the frequency *v* observed when the source is moving, as

$$v = v_0\left(1 + \frac{v_s}{v}\right)^{-1} \qquad (1.49)$$

If $\boldsymbol{v_s}$ is small compared with the wave speed *v* , taking binomial expansion to terms in first order in $\boldsymbol{v_s/v}$ and neglecting higher power , Eq (1.49) may be approximated, giving

$$v = v_0\left(1 - \frac{v_s}{v}\right) \qquad (1.50)$$

For a source approaching the observer, we replace $\boldsymbol{v_s}$ by - $\boldsymbol{v_s}$ to get

$$v = v_0\left(1 + \frac{v_s}{v}\right) \qquad (1.51)$$

The observer thus measures a lower frequency when the source recedes from him than he does when it is at rest. He measures a higher frequency when the source approaches him.

1.8.2 Observer Moving: Source Stationary

Now to derive the Doppler shift when the observer is moving with velocity v_0 towards the source and the source is at rest, we have to proceed in a different manner. We work in the reference frame of the moving observer. In this reference frame the source and medium are approaching at speed v_0 and the speed with which the wave approaches is $v_0 + v$. Following a similar procedure as in the previous case, we find that the time interval between the arrival of the first and the (n + 1)th crests is

$$t_{n+1} - t_1 = n\,T_0 - \frac{nv_0T_0}{v_0\ v}$$

The observer thus, measures the period of the wave to be

$$= T_0\left(1 - \frac{v_0}{v_0 + v}\right)$$

$$= T_0\left(1 + \frac{v_0}{v}\right)^{-1}$$

Giving

$$v = v_0\left(1 + \frac{v_0}{v}\right) \qquad (1.52)$$

If $\frac{v_0}{v}$ is small, the Doppler shift is almost same whether it is the observer or the source moving since Eq.(1.52) and the approximate relation Eq.(1.50) are the same.

1.8.3 Both Source and Observer Moving

We will now derive a general expression for Doppler shift when both the source and the observer are moving. As before, let us take the direction from the observer to the source as the positive direction. Let the source and the observer be moving with velocities v_s and v_0 respectively as shown in Fig 1.18 . Suppose at time t = 0 , the observer is at O_1 and the source is at S_1 , O_1 being to the left of S_1 . The source emits a wave of velocity v , of frequency v and period T_0 all measures by an observer at rest with respect to the medium. Let L be the distance between O_1 and S_1 at t = 0 , when the source emits the first crest. Now , since the observer is moving, the velocity of the wave relative to the observer is $v + v_0$. Therefore , the first crest reaches the observer at time $t_1 = L/(v + v_0)$. At time $t = T_0$, both the observer and the source have moved to their new positions O_2 and S_2 respectively. The new distance between the observer and the source . O_2 S_2 , would be $[L + (v_s - v_0)T_0]$. At S_2 , the source emits a second crest.

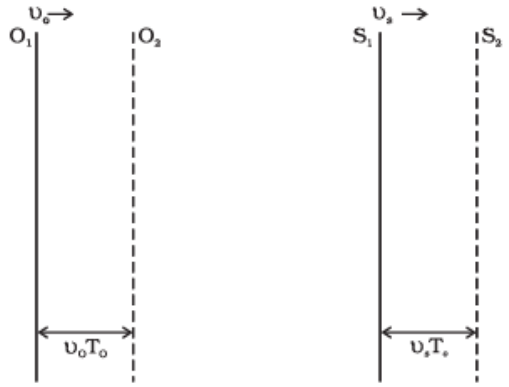


Fig. 1.18 Doppler effect when both the source and observer are moving with different velocities.

This reaches the observer at time

$$t_2 = T_0 + [L + (v_s - v_0)T_0]/\ (v + v_0)$$

At time nT_0 the source emits its $(n+1)$ th crest and this reaches the observer at time

$$t_{n+1} = n\,T_0 + [L + n(v_s - v_0)T_0]/\ (v + v_0)$$

Hence, in a time interval $t_{n+1} - t_1$, i.e.,

$$n\,T_0 + [L + n\,(v_s - v_0)T_0]/\ (v + v_0) - L/(v + v_0),$$

The observer counts n crests and the observer records the period of the wave as equal to T given by

$$T = T_0\left(1 + \frac{v_s - v_0}{v + v_0}\right) = T_0\left(\frac{v + v_s}{v + v_0}\right) \qquad (1.53)$$

The frequency v observer by the observer is given by

$$v = v_0\left(\frac{v + v_0}{v - v_s}\right) \qquad (1.54)$$

Consider a passenger sitting in a train moving on a straight track. Suppose she hears a whistle sounded by the driver of the train . What frequency will she measure or hear ? Here both the observer and the source are moving with the same velocity, so there will be no shift is frequency and the passenger will note the natural frequency. But an observer outside who

is stationary with respect to the track will note a higher frequency if the train is approaching him and a lower frequency when it recedes from him.

Note that we have defined the direction from the observer to the source as the positive direction. Therefore, if the observer is moving towards the source, v_0 has a positive (numerical value whereas if O is moving away from S, v_0 has a negative value. On the other hand, if S is moving away from O, v_s has a positive value whereas if it is moving towards O, v_s has a negative value . The sound emitted by the source travels in all direction. It is that part of sound coming towards the observer which the observer receives and detects. Therefore, the relative velocity of sound with respect to the observer is $\boldsymbol{v + v_0}$ in all cases.

SUMMARY

1. *Mechanical waves* can exist in material media and are governed by Newton's Laws.
2. *Transverse waves* are waves in which the particles of the medium oscillate perpendicular to the direction of wave propagation.
3. *Longitudinal waves* are waves in which the particles of the medium oscillate along the direction of wave propagation.
4. *Progressive wave* is a wave that moves from one point of medium to another.
5. *The displacement* in a sinusoidal wave propagating in the positive x direction is given by
   $y\,(x, t) = a \sin (kx - \omega t + \phi)$
   where *a is the* amplitude of the wave, *k* is the angular wave number, ω *is the* angular frequency, $(kx - \omega t + \phi)$ *is the* phase, *and* ϕ *is the* phase constant or phase angle.
6. *Wavelength* λ of a progressive wave is the distance between two consecutive points of the same phase at a given time. In a stationary wave, it is twice the distance between two consecutive nodes or antinodes.
7. *Period T* of oscillation of a wave is defined as the time any element of the medium takes to move through one complete oscillation. It is related to the *angular frequency* ω through the relation
   $$T = \frac{2\pi}{\omega}$$
8. *Frequency v* of a wave is defined as 1/*T* and is related to angular frequency by
   $$\nu = \frac{\omega}{2\pi}$$
9. *Speed* of a progressive wave is given by $v = \frac{\omega}{k} = \frac{\lambda}{T} = \lambda\nu$
10. *The speed of a transverse wave* on a stretched string is set by the properties of the string. The speed on a string with tension *T* and linear mass density μ is
    $$v = \sqrt{\frac{T}{\mu}}$$
11. *Sound waves* are longitudinal mechanical waves that can travel through solids, liquids, or gases. The speed *v* of sound wave in a fluid having *bulk modulus B* and density ρ is
    $$v = \sqrt{\frac{B}{\rho}}$$
    The speed of longitudinal waves in a metallic bar is

$$v = \sqrt{\frac{Y}{\rho}}$$

For gases, since $B = \gamma P$, the speed of sound is

$$v = \sqrt{\frac{\gamma P}{\rho}}$$

12. When two or more waves traverse simultaneously in the same medium, the displacement of any element of the medium is the algebraic sum of the displacements due to each wave. This is known as the *principle of superposition* of waves

$$y = \sum_{t=1}^{n} f_t(x - vt)$$

13. Two sinusoidal waves on the same string exhibit *interference*, adding or cancelling according to the principle of superposition. If the two are travelling in the same direction and have the same amplitude *a* and frequency but differ in phase by a *phase constant* ϕ, the result is a single wave with the same frequency ω:

$$y(x,t) = \left[2\,a\cos\frac{1}{2}\phi\right]\sin(kx - \omega t + \frac{1}{2}\phi)$$

If $\phi = 0$ or an integral multiple of 2π, the waves are exactly in phase and the interference is constructive; if $\phi = \pi$, they are exactly out of phase and the interference is destructive.

14. A travelling wave, at a rigid boundary or a closed end, is reflected with a phase reversal but the reflection at an open boundary takes place without any phase change.

For an incident wave

$y_i(x, t) = a \sin(kx - \omega t)$

the reflected wave at a rigid boundary is

$y_r(x, t) = -a \sin(kx + \omega t)$

For reflection at an open boundary

$y_r(x,t) = a \sin(kx + \omega t)$

15. The interference of two identical waves moving in opposite directions produces *standing waves.* For a string with fixed ends, the standing wave is given by

$y(x, t) = [2a \sin kx] \cos \omega t$

Standing waves are characterised by fixed locations of zero displacement called *nodes* and fixed locations of maximum displacements called *antinodes.* The separation between two consecutive nodes or antinodes is $\lambda/2$.

A stretched string of length *L* fixed at both the ends vibrates with frequencies given by $\nu = n\frac{v}{2L}$ $n = 1, 2, 3, ...$

The set of frequencies given by the above relation are called the *normal modes* of oscillation of the system. The oscillation mode with lowest frequency is called the *fundamental mode* or the *first harmonic.* The *second harmonic* is the oscillation mode with $n = 2$ and so on.

A pipe of length L with one end closed and other end open (such as air columns) vibrates with frequencies given by

$$\nu = \left(n + \frac{1}{2}\right)\frac{v}{2L}, \; n = 0, 1, 2, 3, \ldots$$

The set of frequencies represented by the above relation are the *normal modes* of oscillation of such a system. The lowest frequency given by $v/4L$ is the fundamental mode or the first harmonic.

16. A string of length L fixed at both ends or an air column closed at one end and open at the other end or open at both the ends, vibrates with certain frequencies called their normal modes. Each of these frequencies is a *resonant frequency* of the system.
17. *Beats* arise when two waves having slightly different frequencies, ν_1 and ν_2 and comparable amplitudes, are superposed. The beat frequency is

    $\nu_{beat} = \nu_1 \sim \nu_2$
18. The *Doppler effect* is a change in the observed frequency of a wave when the source (S) or the observer (O) or both move(s) relative to the medium. For sound the observed frequency ν is given in terms of the source frequency ν_o by

$$\nu = \nu_o\left(\frac{v + v_0}{v + v_S}\right)$$

here v is the speed of sound through the medium, v_0 is the velocity of observer relative to the medium, and v_S is the source velocity relative to the medium. In using this formula, velocities in the direction OS should be treated as positive and those opposite to it should be taken to be negative.

VERY SHORT ANSWER QUESTIONS (2 MARKS)

1. What does a wave represent?
2. Distinguish between transverse and longitudinal waves.
3. Obtain an expression for the wave velocity in terms of these parameters.
4. What is the principle of superposition of waves?
5. Under what conditions will a wave be reflected?
6. What is the principle of superposition of waves?
7. Under what conditions will a wave be reflected?
8. What is a stationary or standing wave?
9. What do you understand by the terms ‘node’ and ‘antinode’?
10. What are harmonics?
11. What are ‘beats’?
12. What is ‘Doppler effect’ ? Give an example.

SHORT ANSWER QUESTIONS (4 MARKS)

1. What are transverse waves ? Give illustrative examples of such waves ?
2. What are longitudinal waves ? Give illustrative examples of such waves.
3. Write an expression for a progressive harmonic wave and explain the various parameters used in the expression.
4. Explain the modes of vibration of a stretched string with examples.
5. Explain the modes of vibration of an air column in an open pipe.
6. What are beats ? When do they occur ? Explain their use, if any.
7. What is Doppler effect ? Give illustrative examples.

LONG ANSWER QUESTIONS (8 MARKS)

1. Explain the formation of stationary waves in stretched strings and hence deduce the laws of transverse waves in stretched strings.
2. Explain the formation of stationary waves in an air column enclosed in open pipe. Derive the equations for the frequencies of the harmonic produced.
3. How are stationary waves formed in closed pipes? Explain the various modes of vibrations and obtained relations for their frequencies.
4. What are the beats ? Obtained an expression for the beat frequency .Where and how are beats made use of?
5. What is Doppler effect? Obtained an expression for the apparent frequency of sound heard when the source is in motion with respect to an observer at rest.
6. What is Doppler shift ? Obtained an expression for the apparent frequency of sound heard when the observer is in motion with respect to a source at rest.

CHAPTER 2

RAY OPTICS AND OPTICAL INSTRUMENTS

2.1 INTRODUCTION

Nature has endowed the human eye (retina) with the sensitivity to detect electromagnetic waves within a small range of the electromagnetic spectrum. Electromagnetic radiation belonging to this region of the spectrum (wavelength of about 400 nm to 750 nm) is called light. It is mainly through light and the sense of vision that we know and interpret the world around us.

There are two things that we can intuitively mention about light from common experience. First, that it travels with enormous speed and second, that it travels in a straight line. It took some time for people to realise that the speed of light is finite and measurable. Its presently accepted value in vacuum is $c = 2.99792458 \times 10^8$ m s^{-1}. For many purposes, it suffices to take $c = 3 \times 10^8$ ms^{-1}. The speed of light in vacuum is the highest speed attainable in nature.

The intuitive notion that light travels in a straight line seems to contradict what we have learnt in Chapter 8, that light is an electromagnetic wave of wavelength belonging to the visible part of the spectrum. How to reconcile the two facts? The answer is that the wavelength of light is very small compared to the size of ordinary objects that we encounter commonly (generally of the order of a few cm or larger). In this situation, as you will learn in Chapter 10, a light wave can be considered to travel from one point to another, along a straight line joining them. The path is called a ray of light, and a bundle of such rays constitutes a beam of light.

In this chapter, we consider the phenomena of reflection, refraction and dispersion of light, using the ray picture of light. Using the basic laws of reflection and refraction, we shall study the image formation by plane and spherical reflecting and refracting surfaces. We then go on to describe the construction and working of some important optical instruments, including the human eye.

PARTICLE MODEL OF LIGHT

Newton's fundamental contributions to mathematics, mechanics, and gravitation often blind us to his deep experimental and theoretical study of light. He made pioneering contributions in field of optics. He further developed the corpuscular model of light proposed by Descartes. It presumes that light energy is concentrated in tiny particles called *corpuscles.* He further assumed that corpuscles of light were mass less elastic particles. With his understanding of mechanics, he could come up with a simple model of reflection and refraction. It is a common observation that a ball bouncing from a smooth plane surface obeys the laws of reflection. When this is an elastic collision, the magnitude of the velocity remains the same. As the surface is smooth, there is no force acting parallel to the surface, so the component of momentum in this direction also remains the same. Only the component perpendicular to the surface, i.e., the normal component of the momentum, gets reversed in reflection. Newton argued that smooth surfaces like mirrors reflect the corpuscles in a similar manner.

In order to explain the phenomena of refraction, Newton postulated that the speed of the corpuscles was greater in water or glass than in air. However, later on it was discovered that the speed of light is less in water or glass than in air.

In the field of optics, Newton-the experiment, was greater than Newton-the theorist. He himself observed many phenomena, which were difficult to understand in terms of particle

nature of light. For example, the colors observed due to understand in terms of particle nature of light for example, the colours observed due to a thin film of oil on water. Property of partial reflection of light is yet another such example. Everyone who has looked into the water in a pond sees image of the face in it, but also sees the bottom of the pond. Newton argued that some of the corpuscles, which fall on the water, get reflected and some get transmitted. But what property could distinguish these two kinds of corpuscles? Newton had to postulate some kind of unpredictable, chance phenomenon, which decided whether an individual corpuscle would be reflected or not. In explaining other phenomena, however, the corpuscles were presumed to behave as if they are identical. Such a dilemma does not occur in the wave picture of light an incoming wave can be divided into two weaker waves at the boundary between air and water.

2.2 REFLECTION OF LIGHT BY SPHERICAL MIRRORS

We are familiar with the laws of reflection. The angle of reflection (i.e., the angle between reflected ray and the normal to the reflecting surface or the mirror) equals the angle of incidence (angle between incident ray and the normal). Also that the incident ray, reflected ray and the normal to the reflecting surface at the point of incidence lie in the same plane (Fig. 2.1). These laws are valid at each point on any reflecting surface whether plane or curved. However, we shall restrict our discussion to the special case of curved surfaces, that is, spherical surfaces. The normal in this case is to be taken as normal to the tangent surface at the point of incidence. That is, the normal is along the radius, the line joining the centre of curvature of the mirror to the point of incidence.

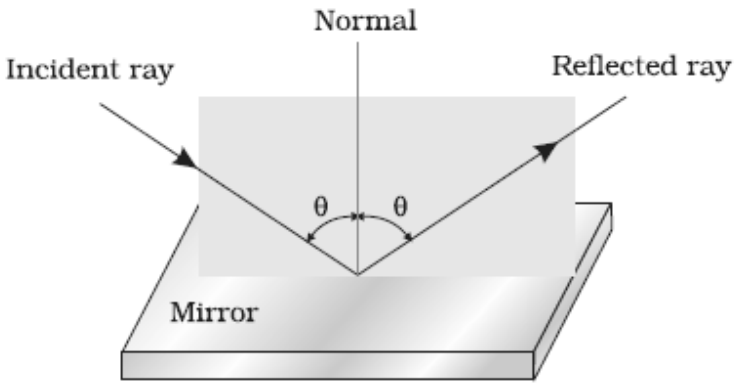


Fig. 2.1 The incident ray, reflected ray and the normal to the reflecting surface lie in the same plane.

We have already studied that the geometric centre of a spherical mirror is called its pole while that of a spherical lens is called its optical centre. The line joining the pole and the centre of curvature of the spherical mirror is known as the principal axis. In the case of spherical lenses, the principal axis is the line joining the optical centre with its principal focus as you will see later.

2.2.1 Sign convention

To derive the relevant formulae for reflection by spherical mirrors and refraction by spherical lenses, we must first adopt a sign convention for measuring distances. In this book, we shall follow the Cartesian sign convention. According to this convention, all distances are measured from the pole of the mirror or the optical centre of the lens. The distances measured in the same direction as the incident light are taken as positive and those measured in the direction

opposite to the direction of incident light are taken as negative (Fig. 2.2). The heights measured upwards with respect to x-axis and normal to the principal axis (x-axis) of the mirror/ lens are taken as positive (Fig. 2.2). The heights measured downwards are taken as negative. With a common accepted convention, it turns out that a single formula for spherical mirrors and a single formula for spherical lenses can handle all different cases.

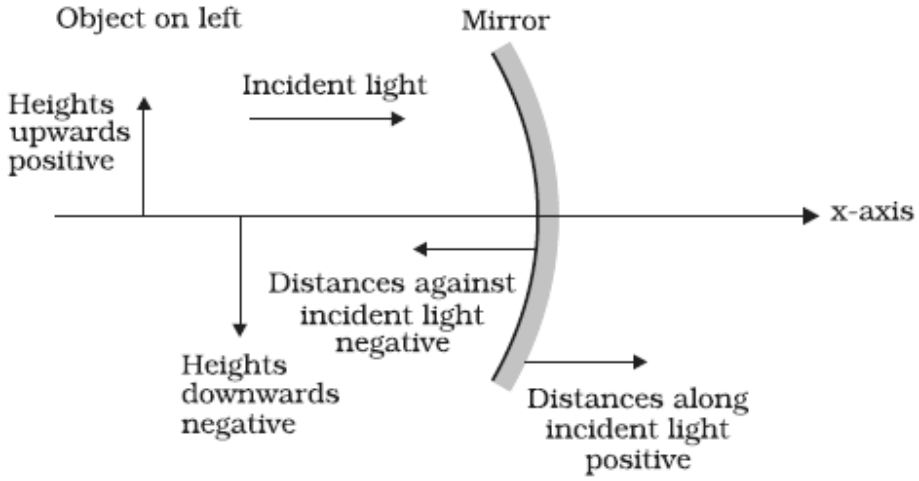


Fig. 2.2 The Cartesian Sign convention

2.2.2 Focal length of spherical mirrors

Figure 2.3 shows what happens when a parallel beam of light is incident on (a) a concave mirror, and (b) a convex mirror. We assume that the rays are paraxial, i.e., they are incident at points close to the pole P of the mirror and make small angles with the principal axis. The reflected rays converge at a point F on the principal axis of a concave mirror [Fig. 2.3(a)]. For a convex mirror, the reflected rays appear to diverge from a point F on its principal axis [Fig. 2.3(b)]. The point F is called the principal focus of the mirror. If the parallel paraxial beam of light were incident, making some angle with the principal axis, the reflected rays would converge (or appear to diverge) from a point in a plane through F normal to the principal axis. This is called the focal plane of the mirror [Fig. 2.3(c)].

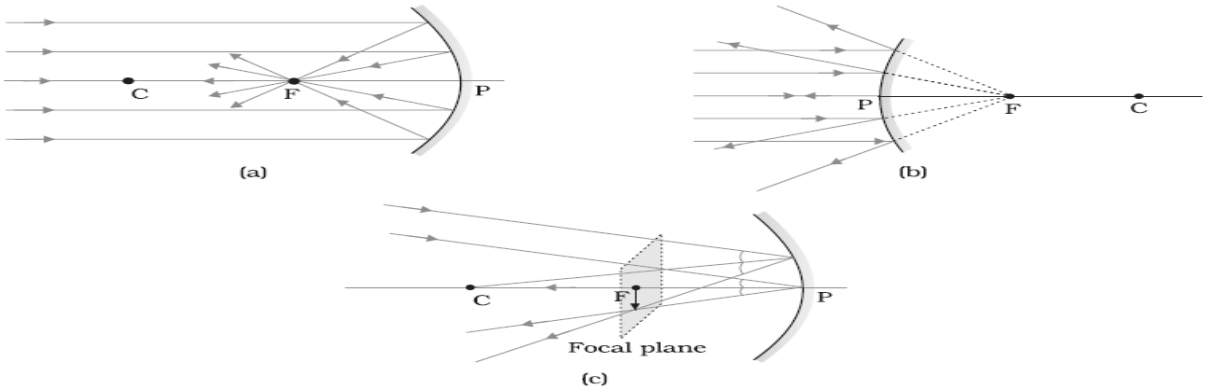


Fig. 2.3 Focus of a concave and convex mirror.

The distance between the focus F and the pole P of the mirror is called the focal length of the mirror, denoted by f. We now show that f = R/2, where R is the radius of curvature of the mirror. The geometry of reflection of an incident ray is shown in Fig. 2.4.

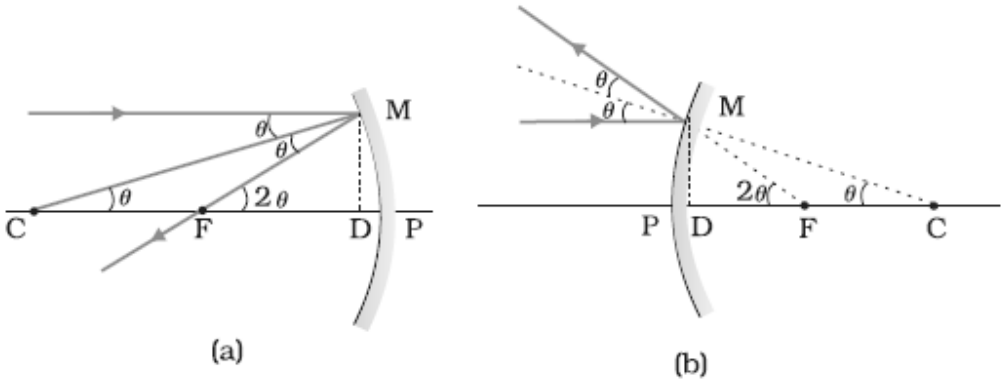


Fig. 9.4 Geometry of reflection of an incident ray on (a) concave spherical mirror, and (b) convex spherical mirror.

Let C be the centre of curvature of the mirror. Consider a ray parallel to the principal axis striking the mirror at M. Then CM will be perpendicular to the mirror at M. Let θ be the angle of incidence, and MD be the perpendicular from M on the principal axis. Then,

∠MCP = θ and ∠MFP = 2θ

Now,

$\tan\theta = \frac{MD}{CD}$ and $\tan\theta = \frac{MD}{FD}$ (2.1)

For small θ, which is true for paraxial rays, tanθ ≈ θ, tan2θ ≈ 2θ. Therefore, Eq. (2.1) gives

$\frac{MD}{FD} = 2\frac{MD}{CD}$

Or, $FD = CD/2$ (2.2)

Now, for small θ, the point D is very close to the point P.

Therefore, FD = f and CD = R. Equation (2.2) then gives

f = R/2 (2.3)

2.2.3 The mirror equation

If rays emanating from a point actually meet at another point after reflection and/or refraction, that point is called the ***image*** of the first point. The image is ***real*** if the rays actually converge to the point; it is ***virtual*** if the rays do not actually meet but appear to diverge from the point when produced backwards. An image is thus a point-to-point correspondence with the object established through reflection and/or refraction.

In principle, we can take any two rays emanating from a point on an object, trace their paths, find their point of intersection and thus, obtain the image of the point due to reflection at a spherical mirror. In practice, however, it is convenient to choose any two of the following rays:

(i) The ray from the point which is parallel to the principal axis. The reflected ray goes through the focus of the mirror.

(ii) The ray passing through the centre of curvature of a concave mirror or appearing to pass through it for a convex mirror. The reflected ray simply retraces the path.

(iii) The ray passing through (or directed towards) the focus of the concave mirror or appearing to pass through (or directed towards) the focus of a convex mirror. The reflected ray is parallel to the principal axis.

(iv) The ray incident at any angle at the pole. The reflected ray follows laws of reflection. Figure 2.5 shows the ray diagram considering three rays. It shows the image A′B′ (in this case, real) of an object AB formed by a concave mirror. It does not mean that only three rays emanate from the point A. An infinite number of rays emanate from any source, in all directions. Thus, point A′ is image point of A if every ray originating at point A and falling on the concave mirror after reflection passes through the point A′. We now derive the mirror equation or the relation between the object distance (u), image distance (v) and the focal length (f).

From Fig. 2.5, the two right-angled triangles A′B′F and MPF are similar. (For paraxial rays, MP can be considered to be a straight line perpendicular to CP.) Therefore,

$$\frac{B'A'}{PM} = \frac{B'F}{FP} \text{ or } \frac{B'A'}{BA} = \frac{B'F}{FP} \text{ or } = (\because PM = AB) \quad (2.4)$$

Since ∠APB = ∠A′PB′, the right angled triangles A′B′P and ABP are also similar. Therefore,

$$\frac{B'A'}{BA} = \frac{B'P}{BP} \quad (2.5)$$

Comparing Eqs. (2.4) and (2.5), we get

$$\frac{B'F}{FP} = \frac{B'P-FP}{FP} = \frac{B'P}{BP} \quad (2.6)$$

Equation (2.6) is a relation involving magnitude of distances. We now apply the sign convention. We note that light travels from the object to the mirror MPN. Hence this is taken as the positive direction. To reach the object AB, image A′B′ as well as the focus F from the pole P, we have to travel opposite to the direction of incident light. Hence, all the three will have negative signs. Thus,

$$B'P = -v, FP = -f, BP = -u$$

Using these in Eq. (2.6), we get

$$\frac{-v+f}{-f} = \frac{-v}{-u}$$

Or $$\frac{v-f}{f} = \frac{v}{u}$$

$$\frac{v}{f} = 1 + \frac{v}{u}$$

Dividing it by v, we get

$$\frac{1}{v} + \frac{1}{u} = \frac{1}{f} \quad (2.7)$$

This relation is known as the ***mirror equation***.

The size of the image relative to the size of the object is another important quantity to consider. We define linear *magnification (m)* as the ratio of the height of the image (h′) to the height of the object *(h)*:

$$m = \frac{h'}{h} \quad (2.8)$$

h and h' will be taken positive or negative in accordance with the accepted sign convention. In triangles A′B′P and ABP, we have,

$$\frac{B'A'}{BA} = \frac{B'P}{BP}$$

With the sign convention, this becomes

$$\frac{-h'}{h} = \frac{-v}{-u}$$

so that $m = \frac{h'}{h} = -\frac{v}{u}$ (2.9)

We have derived here the mirror equation, Eq. (2.7), and the magnification formula, Eq. (2.9), for the case of real, inverted image formed by a concave mirror. With the proper use of sign convention, these are, in fact, valid for all the cases of reflection by a spherical mirror (concave or convex) whether the image formed is real or virtual. Figure 2.6 shows the ray diagrams for virtual image formed by a concave and convex mirror. You should verify that Eqs. (2.7) and (2.9) are valid for these cases as well.

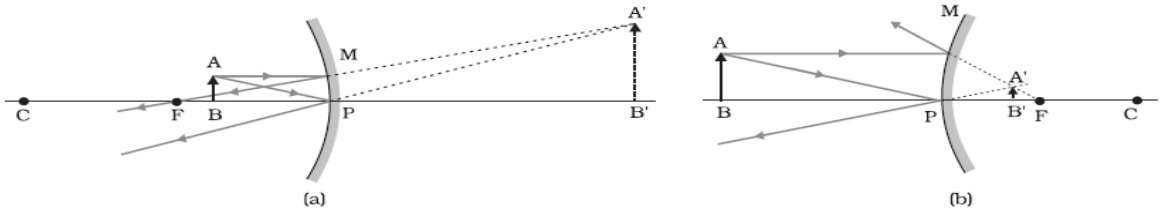


Fig.2.6 Image formation by (a) a concave mirror with object between P and F, and (b) a convex mirror.

2.3 REFRACTION

When a beam of light encounters another transparent medium, a part of light gets reflected back into the first medium while the rest enters the other. A ray of light represents a beam. The direction of propagation of an obliquely incident ($0° < i < 90°$) ray of light that enters the other medium, changes at the interface of the two media. This phenomenon is called refraction of light. Snell experimentally obtained the following laws of refraction:

(i) The incident ray, the refracted ray and the normal to the interface at the point of incidence, all lie in the same plane.

(ii) The ratio of the sine of the angle of incidence to the sine of angle of refraction is constant. Remember that the angles of incidence *(i)* and refraction *(r)* are the angles that the incident and its refracted ray make with the normal, respectively. We have

$$\frac{\sin i}{\sin r} = n_{21} \quad (2.10)$$

where n_{21} is a constant, called the refractive index of the second medium with respect to the first medium. Equation (2.10) is the well-known Snell's law of refraction. We note that n_{21} is a characteristic of the pair of media (and also depends on the wavelength of light), but is independent of the angle of incidence.

From Eq. (2.10), if $n_{21} > 1$, r < i, i.e., the refracted ray bends towards the normal. In such a case medium 2 is said to be optically denser (or denser, in short) than medium 1. On the other hand, if $n_{21} < 1$, r > i, the refracted ray bends away from the normal. This is the case when incident ray in a denser medium refracts into a rarer medium.

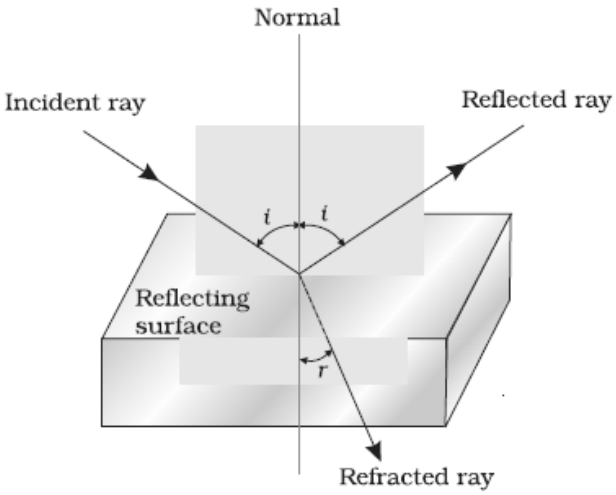


Fig. 2.7 Refraction and reflection of light.

Note: *Optical density should not be confused with mass density, which is mass per unit volume. It is possible that mass density of an optically denser medium may be less than that of an optically rarer medium (optical density is the ratio of the speed of light in two media). For example, turpentine and water. Mass density of turpentine is less than that of water but its optical density is higher.*

If n_{21} is the refractive index of medium 2 with respect to medium 1 and n_{12} the refractive index of medium 1 with respect to medium 2, then it should be clear that

$$n_{12} = \frac{1}{n_{21}} \tag{2.11}$$

It also follows that if n_{32} is the refractive index of medium 3 with respect to medium 2 then $n_{32} = \times n_{12}$, where n_{31} is the refractiven31index of medium 3 with respect to medium 1. Some elementary results based on the laws of refraction follow immediately. For a rectangular slab, refraction takes place at two interfaces (air-glass and glass-air). It is easily seen from Fig. 2.8 that $r_2 = i_1$, i.e., the emergent ray is parallel to the incident ray—there is no deviation, but it does suffer lateral displacement/ shift with respect to the incident ray.

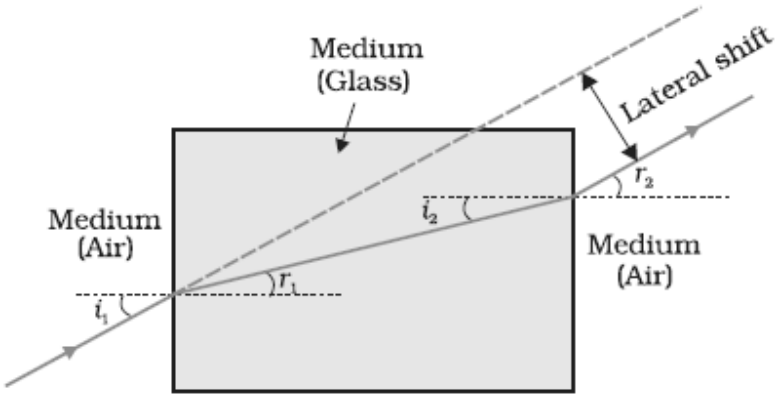


Fig. 2.8 Lateral shift of a ray refracted through a parallel-sided slab.

Another familiar observation is that the bottom of a tank filled with water appears to be raised (Fig. 2.9). For viewing near the normal direction, it can be shown that the apparent depth, (h_1) is real depth (h_2) divided by the refractive index of the medium (water).

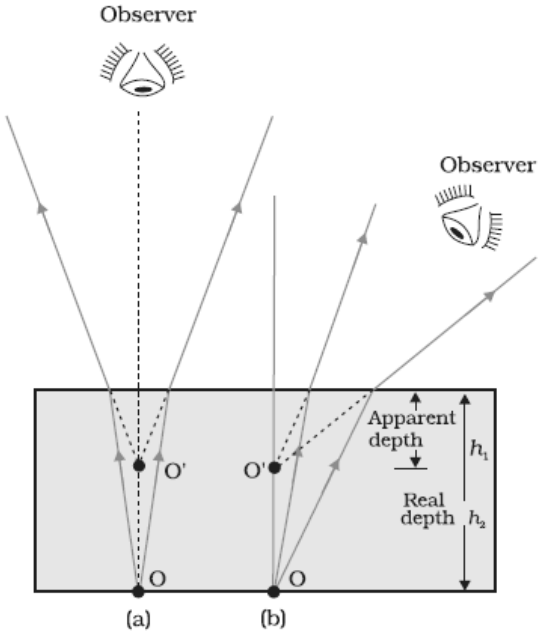


Fig. 2.9 Apparent depth for (a) normal, and (b) oblique viewing.

The refraction of light through the atmosphere is responsible for many interesting phenomena. For example, the sun is visible a little before the actual sunrise and until a little after the actual sunset due to refraction of light through the atmosphere (Fig. 2.10). By actual sunrise we mean the actual crossing of the horizon by the sun. Figure 2.10 shows the actual and apparent positions of the sun with respect to the horizon. The figure is highly exaggerated to show the effect. The refractive index of air with respect to vacuum is 1.00029. Due to this, the apparent shift in the direction of the sun is by about half a degree and the corresponding time difference between actual sunset and apparent sunset is about 2 minutes (see Example 2.5). The apparent flattening (oval shape) of the sun at sunset and sunrise is also due to the same phenomenon.

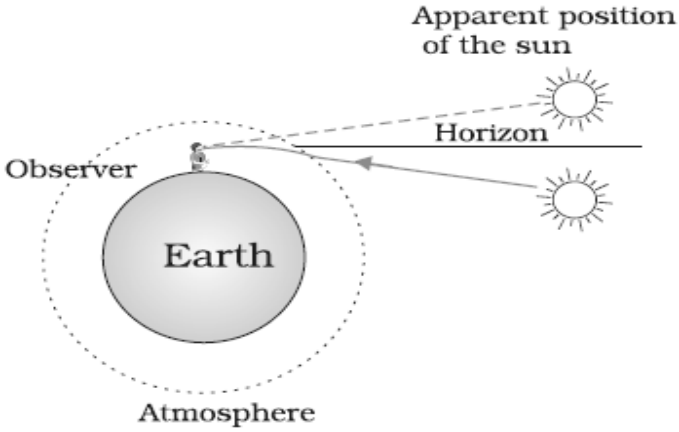


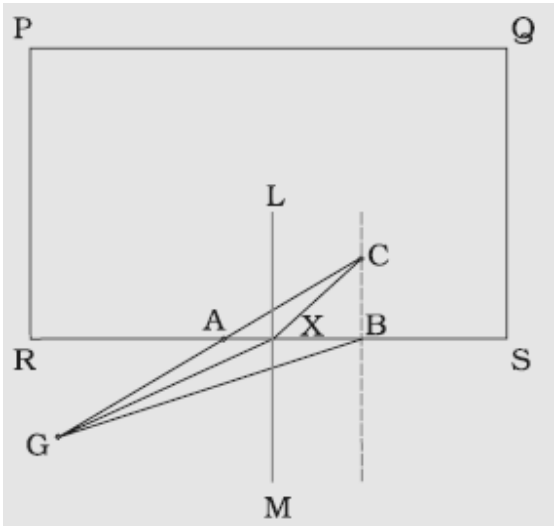
Fig. 2.10 Advance sunrise and delayed sunset due to atmospheric refraction.

THE DROWNING CHILD, LIFEGUARD AND SNELL'S LAW

Consider a rectangular swimming pool PQSR; see figure here. A lifeguard sitting at G outside the pool notices a child drowning at a point C. The guard wants to reach the child in the shortest possible time. Let SR be the side of the pool between G and C. should he/ she take a straight line path GAC between G and C or GBC in which the path BC in water would be the shortest, or some other path GXC? The guard knows that his/her swimming speed v_2.

Suppose the guard enters water at X. Let GX=l_1 and XC=l_2. Then the time taken to reach from G to C would be

$$t = \frac{l_1}{v_1} + \frac{l_2}{v_2}$$



To make this time minimum, one has to differentiate it (with respect to the coordinate of X) and find the point X when t is a minimum. On doing all this algebra (which we skip here), we find that the guard should enter water at a point where Snell's law is satisfied. To understand this, draw a perpendicular LM to side SR at X. Let ∠GXM = i and ∠CXL = r. Then it can be seen that t

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2}$$

In the case of light v_1/v_2, the ration of the velocity of light in vacuum to that in the medium, is the refractive index n of the medium.

In short, whether it is a wave or a particle or a human being, whenever two mediums and two velocities are involved, one must follow Snell's law if one wants to take the shortest time.

2.4 TOTAL INTERNAL REFLECTION

When light travels from an optically denser medium to a rarer medium at the interface, it is partly reflected back into the same medium and partly refracted to the second medium. This reflection is called the *internal reflection.*

When a ray of light enters from a denser medium to a rarer medium, it bends away from the normal, for example, the ray AO^1 B in Fig. 2.11. The incident ray AO_1 is partially reflected (O_1C) and partially transmitted (O_1B) or refracted, the angle of refraction (r) being larger than

the angle of incidence (i). As the angle of incidence increases, so does the angle of refraction, till for the ray AO_3, the angle of refraction is $\pi/2$. The refracted ray is bent so much away from the normal that it grazes the surface at the interface between the two media. This is shown by the ray AO_3D in Fig. 2.11. If the angle of incidence is increased still further (e.g., the ray AO_4), refraction is not possible, and the incident ray is totally reflected.

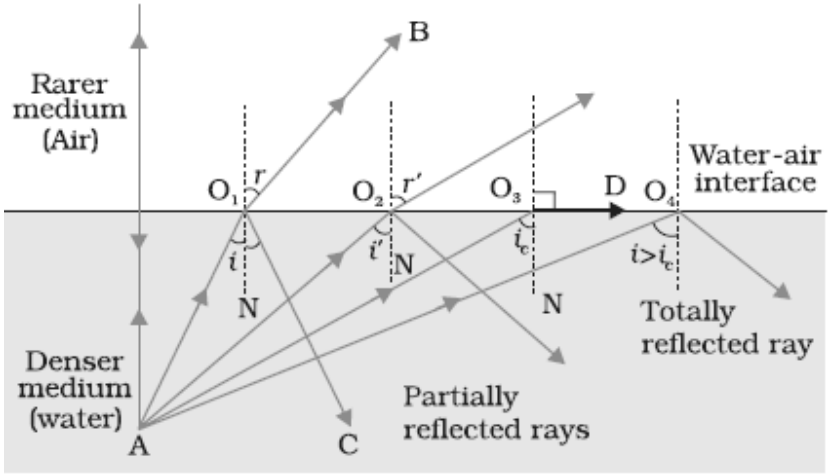


Fig. 2.11 Refraction and internal reflection of rays from a point A in the denser medium (water) incident at different angles at the interface with a rarer medium (air).

This is called *total internal reflection*. When light gets reflected by a surface, normally some fraction of it gets transmitted. The reflected ray, therefore, is always less intense than the incident ray, howsoever smooth the reflecting surface may be. In total internal reflection, on the other hand, no transmission of light takes place.

The angle of incidence corresponding to an angle of refraction 90º, say ∠AO3N, is called the critical angle (ic) for the given pair of media. We see from Snell's law [Eq. (2.10)] that if the relative refractive index is less than one then, since the maximum value of sin r is unity, there is an upper limit to the value of sin i for which the law can be satisfied, that is, $i = i_c$ such that

$$\sin i_c = n_{21} \qquad (2.12)$$

For values of i larger than i_c, Snell's law of refraction cannot be satisfied, and hence no refraction is possible.

The refractive index of denser medium 2 with respect to rarer medium 1 will be $n_{12} = 1/\sin i_c$. Some typical critical angles are listed in Table 2.1.

TABLE 2.1 CRITICAL ANGLE OF SOME TRANSPARENT MEDIA

Substance medium	Refractive index	Critical angle
Water	1.33	48.75º
Crown glass	1.52	41.14º
Dense flint glass	1.62	37.31º
Diamond	2.42	24.41º

A demonstration for total internal reflection

All optical phenomena can be demonstrated very easily with the use of a laser torch or pointer, which is easily available nowadays. Take a glass beaker with clear water in it. Stir the water a few times with a piece of soap, so that it becomes a little turbid. Take a laser pointer and shine its beam through the turbid water. You will find that the path of the beam inside the water shines brightly.

Shine the beam from below the beaker such that it strikes at the upper water surface at the other end. Do you find that it undergoes partial reflection (which is seen as a spot on the table below) and partial refraction [which comes out in the air and is seen as a spot on the roof; Fig. 2.12(a)]? Now direct the laser beam from one side of the beaker such that it strikes the upper surface of water more obliquely [Fig. 2.12(b)]. Adjust the direction of laser beam until you find the angle for which the refraction above the water surface is totally absent and the beam is totally reflected back to water. This is total internal reflection at its simplest. Pour this water in a long test tube and shine the laser light from top, as shown in Fig. 2.12(c). Adjust the direction of the laser beam such that it is totally internally reflected every time it strikes the walls of the tube. This is similar to what happens in optical fibres.

Take care not to look into the laser beam directly and not to point it at anybody's face.

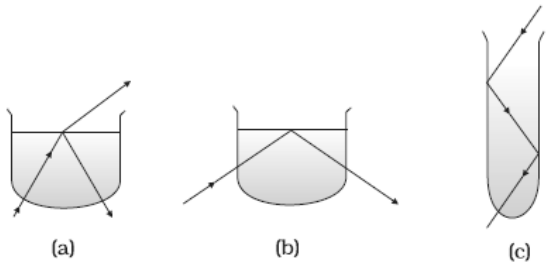


Fig. 2.13 Observing total internal reflection in water with a laser beam (refraction due to glass of beaker neglected being very thin).

2.4.1 Total internal reflection in nature and its technological applications

(i) Mirage: On hot summer days, the air near the ground becomes hotter than the air at higher levels. The refractive index of air increases with its density. Hotter air is less dense, and has smaller refractive index than the cooler air. If the air currents are small, that is, the air is still, the optical density at different layers of air increases with height. As a result, light from a tall object such as a tree, passes through a medium whose refractive index decreases towards the ground. Thus, a ray of light from such an object successively bends away from the normal and undergoes total internal reflection, if the angle of incidence for the air near the ground exceeds the critical angle. This is shown in Fig. 2.13(b). To a distant observer, the light appears to be coming from somewhere below the ground. The observer naturally assumes that light is being reflected from the ground, say, by a pool of water near the tall object. Such inverted images of distant tall objects cause an optical illusion to the observer. This phenomenon is called mirage. This type of mirage is especially common in hot deserts. Some of you might have noticed that while moving in a bus or a car during a hot summer day, a distant patch of road, especially on a highway, appears to be wet. But, you do not find any evidence of wetness when you reach that spot. This is also due to mirage.

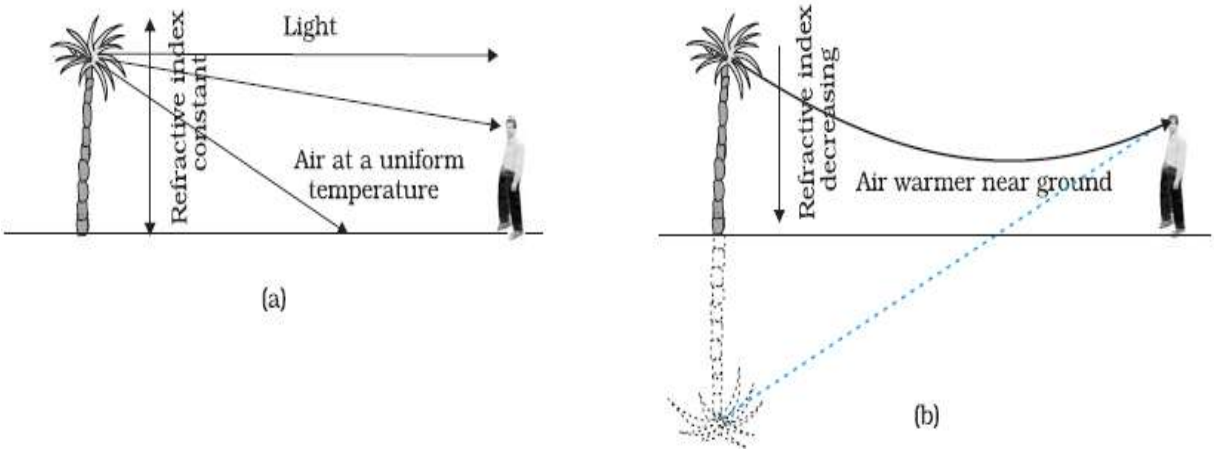


Fig. 2.13 (a) A tree is seen by an observer at its place when the air above the ground is at uniform temperature, (b) When the layers of air close to the ground have varying temperature with hottest layers near the ground, light from a distant tree may undergo total internal reflection, and the apparent image of the tree may create an illusion to the observer that the tree is near a pool of water.

(ii) Diamond: Diamonds are known for their spectacular brilliance. Their brilliance is mainly due to the total internal reflection of light inside them. The critical angle for diamond-air interface (≅ 24.4°) is very small, therefore once light enters a diamond, it is very likely to undergo total internal reflection inside it. Diamonds found in nature rarely exhibit the brilliance for which they are known. It is the technical skill of a diamond cutter which makes diamonds to sparkle so brilliantly. By cutting the diamond suitably, multiple total internal reflections can be made to occur.

(iii)Prism: Prisms designed to bend light by 90º or by 180º make use of total internal 90º and 180º make use of total internal reflection [Fig. 2.14(a) and (b)]. Such a prism is also used to invert images without changing their size [Fig. 2.14(c)]. In the first two cases, the critical angle ic for the material of the prism must be less than 45º. We see from Table 2.1 that this is true for both crown glass and dense flint glass.

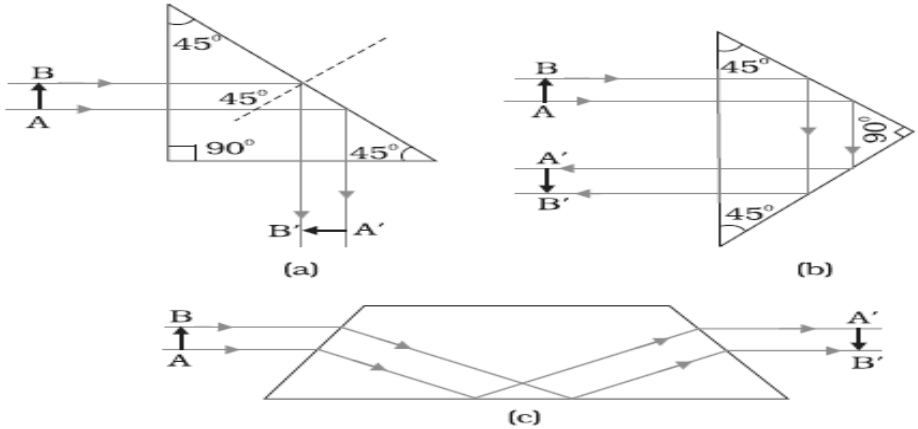


Fig. 2.14 Prisms designed to bend rays by 90º or by 180º or to invert image without changing its size make use of total internal reflection

(iv) Optical fibres: Now-a-days optical fibres are extensively used for transmitting audio and video signals through long distances. Optical fibres too make use of the phenomenon of total internal reflection. Optical fibres are fabricated with high quality composite glass/quartz fibres. Each fibre consists of a core and cladding. The refractive index of the material of the core is higher than that of the cladding.

When a signal in the form of light is directed at one end of the fibre at a suitable angle, it undergoes repeated total internal reflections along the length of the fibre and finally comes out at the other end (Fig. 2.15). Since light undergoes total internal reflection at each stage, there is no appreciable loss in the intensity of the light signal. Optical fibres are fabricated such that light reflected at one side of inner surface strikes the other at an ngle larger than the critical angle. Even if the fibre is bent, light can easily travel along its length. Thus, an optical fibre can be used to act as an optical pipe.

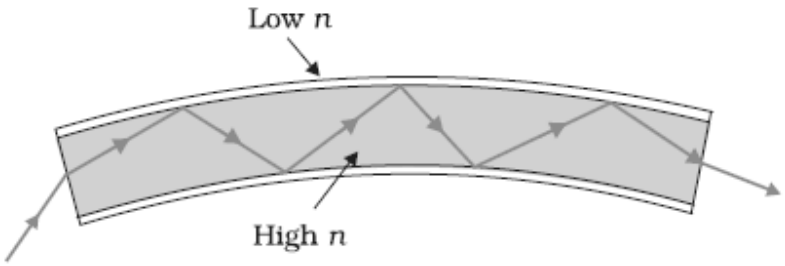


Fig. 2.16 Light undergoes successive total internal reflections as it moves through an optical fibre.

A bundle of optical fibres can be put to several uses. Optical fibres are extensively used for transmitting and receiving electrical signals which are converted to light by suitable transducers. Obviously, optical fibres can also be used for transmission of optical signals. For example, these are used as a 'light pipe' to facilitate visual examination of internal organs like esophagus, stomach and intestines. You might have seen a commonly available decorative lamp with fine plastic fibres with their free ends forming a fountain like structure. The other end of the fibres is fixed over an electric lamp. When the lamp is switched on, the light travels from the bottom of each fibre and appears at the tip of its free end as a dot of light. The fibres in such decorative lamps are optical fibres.

The main requirement in fabricating optical fibres is that there should be very little absorption of light as it travels for long distances inside them. This has been achieved by purification and special preparation of materials such as quartz. In silica glass fibres, it is possible to transmit more than 95% of the light over a fibre length of 1 km. (Compare with what you expect for a block of ordinary window glass 1 km thick.)

2.5 REFRACTION AT SPHERICAL SURFACES AND BY LENSES

We have so far considered refraction at a plane interface. We shall now consider refraction at a spherical interface between two transparent media. An infinitesimal part of a spherical surface can be regarded as planar and the same laws of refraction can be applied at every point on the surface. Just as for reflection by a spherical mirror, the normal at the point of incidence is perpendicular to the tangent plane to the spherical surface at that point and, therefore, passes through its centre of curvature. We first consider refraction by a single spherical surface and follow it by thin lenses. A thin lens is a transparent optical medium bounded by two surfaces; at least one of which should be spherical. Applying the formula for image formation by

a single spherical surface successively at the two surfaces of a lens, we shall obtain the lens maker's formula and then the lens formula.

2.5.1 Refraction at a spherical surface

Figure 2.16 shows the geometry of formation of image I of an object O on the principal axis of a spherical surface with centre of curvature C, and radius of curvature R. The rays are incident from a medium of refractive index n1, to another of refractive index n_2. As before, we take the aperture (or the lateral size) of the surface to be small compared to other distances involved, so that small angle approximation can be made. In particular, NM will be taken to be nearly equal to the length of the perpendicular from the point N on the principal axis. We have, for small angles,

tan ∠NOM = MN/OM

tan ∠NCM = MN/MC

tan ∠NIM = MN/MI

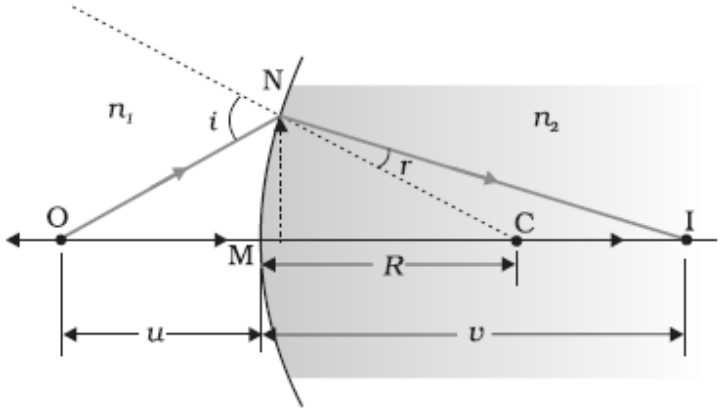


Fig. 2.16 Refraction at a spherical surface separating two media.

LIGHT SOURCES AND PHOTOMETRY

It is known that a body above absolute zero temperature emits electromagnetic radiation. The wavelength region in which the body emits the radiation depends on its absolute temperature. Radiation emitted by a hot body, for example, a tungsten filament lamp having temperature 2850 K are partly invisible and mostly in infrared (or heat) region. As the temperature of the body increases radiation emitted by it is in visible region. The sun with temperature of about 5500 K emits radiation whose energy versus wavelength graph peaks approximately at 550 nm corresponding to green light and is almost in the middle of the visible region. The energy versus wavelength distribution graph for a given body peaks at some wavelength, which is inversely proportional to the absolute temperature of that body.

The measurement of light as perceived by human eye is called photometry. Photometry is measurement of a physiological phenomenon, being the stimulus of light as received by the human eye, transmitted by the optic nerves and analysed by the brain. The main physical quantities in photometry are (i) the luminous intensity of the source,(ii) the luminous flux or flow of light from the source, and (iii) *illuminance* of the surface. The SI unit of luminous intensity (I) is candela (cd). The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency540 × 1012 Hz and that has a radiant intensity in that direction of 1/683 watt per steradian. If a light source emits one candela of luminous intensity

into a solid angle of one steradian, the total luminous flux emitted into that solid angle is one lumen (lm). A standard100 watt incadescent light bulb emits approximately 1700 lumens.

In photometry, the only parameter, which can be measured directly is illuminance. It is defined as luminous flux incident per unit area on a surface (lm/m^2 or lux). Most light meters measure this quantity. The illuminance E, produced by a source of luminous intensity I, is given by E = I/r2, where r is the normal distance of the surface from the source. A quantity named luminance (L), is used to characterise the brightness of emitting or reflecting flat surfaces. Its unit is cd/m^2 (sometimes called 'nit' in industry). A good LCD computer monitor has a brightness of about 250 nits.

Now, for ΔNOC, i is the exterior angle. Therefore, i = ∠NOM + ∠NCM

$$i = \frac{MN}{OM} + \frac{MN}{MC} \tag{2.13}$$

Similarly, r = ∠NCM – ∠NIM

i.e., $$r = \frac{MN}{MC} - \frac{MN}{MI} \tag{2.14}$$

Now, by Snell's law

$$n_1 \sin i = n_2 \sin r$$

or for small angles

$$n_1 i = n_2 r$$

Substituting i and r from Eqs. (2.13) and (2.14), we get

$$\frac{n_1}{OM} + \frac{n_2}{MI} = \frac{n_2 - n_1}{MC} \tag{2.15}$$

Here, OM, MI and MC represent magnitudes of distances. Applying the Cartesian sign convention,

$$OM = -u, MI = +v, MC = +R$$

Substituting these in Eq. (2.15), we get

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R} \tag{2.16}$$

Equation (2.16) gives us a relation between object and image distance in terms of refractive index of the medium and the radius of curvature of the curved spherical surface. It holds for any curved spherical surface.

2.5.2 Refraction by a lens

Figure 2.17(a) shows the geometry of image formation by a double convex lens. The image formation can be seen in terms of two steps:

(i) The first refracting surface forms the image I_1 of the object O [Fig. 2.17(b)]. The image I_1 acts as a virtual object for the second surface that forms the image at I_1 [Fig. 2.17(c)]. Applying Eq. (2.15) to the first interface ABC, we get

$$\frac{n_1}{OB} + \frac{n_2}{BI_1} = \frac{n_2 - n_1}{BC_1} \tag{2.17}$$

A similar procedure applied to the second interface* ADC gives,

$$-\frac{n_{21}}{DI_1} + \frac{n_1}{DI} = \frac{n_2 - n_1}{DC_2} \tag{2.18}$$

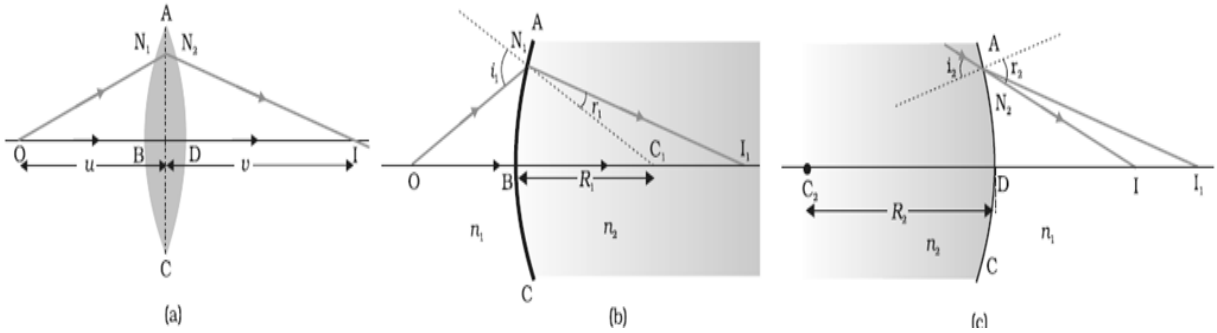


Fig. 2.17 (a) The position of object, and the image formed by a double convex lens (b) Refraction at the first spherical surface and (c) Refraction at the second spherical surface

For a thin lens, $BI_1 = DI_1$. Adding Eqs. (2.17) and (2.18), we get

$$\frac{n_1}{OB} + \frac{n_1}{DI} = \frac{n_1}{f} \tag{2.19}$$

Suppose the object is at infinity, i.e., OB $\rightarrow\infty$ and DI = f, Eq. (2.19) gives

$$\frac{n_1}{f} = (n_2 - n_1)\left(\frac{1}{BC_1} + \frac{1}{DC_2}\right) \tag{2.20}$$

The point where image of an object placed at infinity is formed is called the focus F, of the lens and the distance f gives its focal length. A lens has two foci, F and F′, on either side of it (Fig. 2.18). By the sign convention,

$BC_1 = + R_1$,

$DC_2 = -R_2$

So Eq. (2.20) can be written as

$$\frac{1}{f} = (n_{21} - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \qquad \therefore\ n_{21} = \frac{n_2}{n_1} \tag{2.21}$$

Equation (2.21) is known as the lens maker's formula. It is useful to design lenses of desired focal length using surfaces of suitable radii of curvature. Note that the formula is true for a concave lens also. In that case $R_1$is negative, R_2 positive and therefore, f is negative.

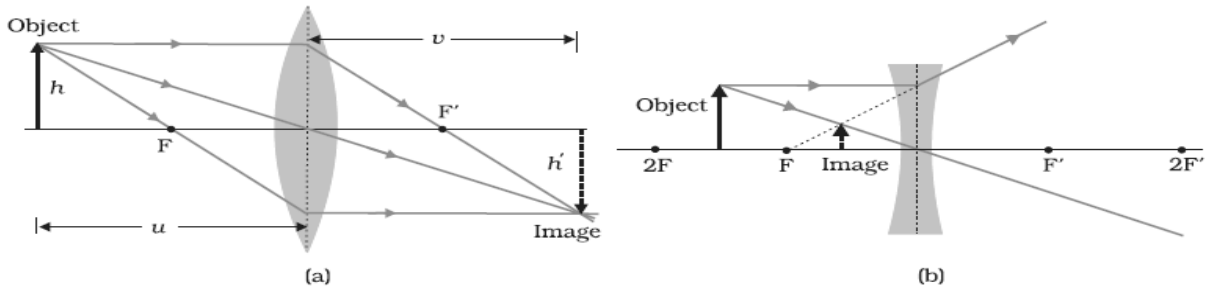


Fig. 2.19 Tracing rays through (a)convex lens (b) concave lens.

---

* Note that now the refractive index of the medium on the right side of ADC is n1 while on its left it is n_2. Further DI_1 is negative as the distance is measured against the direction of incident light.

From Eqs. (2.19) and (2.20), we get

$$\frac{n_1}{OB} + \frac{n_2}{DI} = \frac{n_1}{f} \quad (2.22)$$

Again, in the thin lens approximation, B and D are both close to the optical centre of the lens. Applying the sign convention,

BO = – u, DI = +v, we get

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f} \quad (2.23)$$

Equation (2.23) is the familiar thin lens formula. Though we derived it for a real image formed by a convex lens, the formula is valid for both convex as well as concave lenses and for both real and virtual images. It is worth mentioning that the two foci, F and F′, of a double convex or concave lens are equidistant from the optical centre. The focus on the side of the (original) source of light is called the first focal point, whereas the other is called the *second focal point.*

To find the image of an object by a lens, we can, in principle, take anytwo rays emanating from a point on an object; trace their paths using the laws of refraction and find the point where the refracted rays meet (or appear to meet). In practice, however, it is convenient to choose any two of the following rays:

(i) A ray emanating from the object parallel to the principal axis of the lens after refraction passes through the second principal focus F′ (in a convex lens) or appears to diverge (in a concave lens) from the first principal focus F.

(ii) A ray of light, passing through the optical centre of the lens, emerges without any deviation after refraction.

(iii) A ray of light passing through the first principal focus (for a convex lens) or appearing to meet at it (for a concave lens) emerges parallel to the principal axis after refraction.

Figures 2.18(a) and (b) illustrate these rules for a convex and a concave lens, respectively. You should practice drawing similar ray diagrams for different positions of the object with respect to the lens and also verify that the lens formula, Eq. (2.23), holds good for all cases.

Here again it must be remembered that each point on an object gives out infinite number of rays. All these rays will pass through the same image point after refraction at the lens.

Magnification (m) produced by a lens is defined, like that for a mirror, as the ratio of the size of the image to that of the object. Proceeding in the same way as for spherical mirrors, it is easily seen that for a lens

$$m = \frac{\square'}{} = \frac{v}{u} \quad (2.24)$$

When we apply the sign convention, we see that, for erect (and virtual) image formed by a convex or concave lens, m is positive, while for an inverted (and real) image, m is negative.

2.5.3 Power of a lens

Power of a lens is a measure of the convergence or divergence, which a lens introduces in the light falling on it. Clearly, a lens of shorter focal length bends the incident light more, while converging it in case of a convex lens and diverging it in case of a concave lens. The power P of a lens is defined as the tangent of the angle by which it converges or diverges a beam of light falling at unit distant from the optical centre (Fig. 2.19).

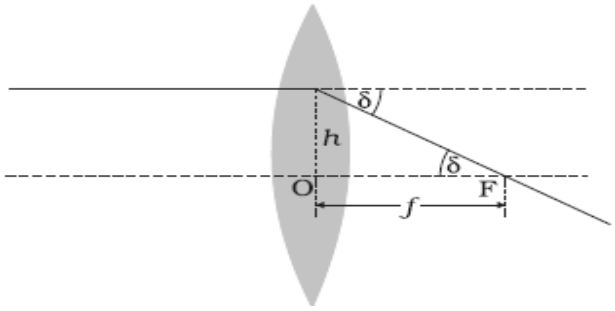


Fig. 2.20 Power of a lens.

$\tan\delta = \frac{h}{f}$; if h=1, $\tan\delta = \frac{1}{f}$ or $\delta = \frac{1}{f}$ for small value of δ. Thus,

$$P = \frac{1}{f} \tag{2.25}$$

The SI unit for power of a lens is dioptre (D): 1D = 1m^{-1}. The power of a lens of focal length of 1 metre is one dioptre. Power of a lens is positive for a converging lens and negative for a diverging lens. Thus, when an optician prescribes a corrective lens of power + 2.5 D, the required lens is a convex lens of focal length + 40 cm. A lens of power of – 4.0 D means a concave lens of focal length – 25 cm.

2.5.4 Combination of thin lenses in contact

Consider two lenses A and B of focal length f_1 and f_2 placed in contact with each other. Let the object be placed at a point O beyond the focus of the first lens A (Fig. 2.20). The first lens produces an image at I_1. Since image I_1 is real, it serves as a virtual object for the second lens B, producing the final image at I. It must, however, be borne in mind that formation of image by the first lens is presumed only to facilitate determination of the position of the final image. In fact, the direction of rays emerging from the first lens gets modified in accordance with the angle at which they strike the second lens. Since the lenses are thin, we assume the optical centres of the lenses to be coincident. Let this central point be denoted by P.

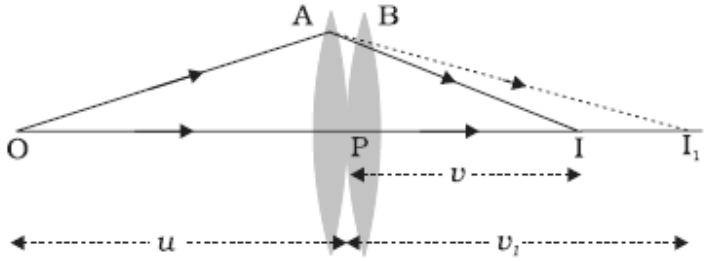


Fig. 2.20 Image formation by a combination of two thin lenses in contact.

For the image formed by the first lens A, we get

$$\frac{1}{v_1} - \frac{1}{u} = \frac{1}{f_1} \tag{2.27}$$

For the image formed by the second lens B, we get

$$\frac{1}{v} - \frac{1}{v_1} = \frac{1}{f_2} \tag{2.28}$$

Adding Eqs. (2.27) and (2.28), we get

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f_1} + \frac{1}{f_2} \quad (2.29)$$

If the two lens-system is regarded as equivalent to a single lens of focal length f, we have

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f}$$

so that we get

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} \quad (2.30)$$

The derivation is valid for any number of thin lenses in contact. If several thin lenses of focal length $f_1, f_2, f_3, \ldots$ are in contact, the effective focal length of their combination is given by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} + \frac{1}{f_3} + \cdots \quad (2.31)$$

In terms of power, Eq. (2.31) can be written as

$$P = P_1 + P_2 + P_3 + \ldots \quad (2.32)$$

where P is the net power of the lens combination. Note that the sum in Eq. (2.32) is an algebraic sum of individual powers, so some of the terms on the right side may be positive (for convex lenses) and some negative (for concave lenses). Combination of lenses helps to obtain diverging or converging lenses of desired magnification. It also enhances sharpness of the image. Since the image formed by the first lens becomes the object for the second, Eq. (2.25) implies that the total magnification mof the combination is a product of magnification (m_1, m_2, m_3,...) of individual lenses

$$m = m_1 m_2 m_3 \ldots \quad (2.33)$$

Such a system of combination of lenses is commonly used in designing lenses for cameras, microscopes, telescopes and other optical instruments.

2.6 REFRACTION THROUGH A PRISM

Figure 2.22 shows the passage of light through a triangular prism ABC. The angles of incidence and refraction at the first face AB are i and r_1, while the angle of incidence (from glass to air) at the second face AC is r_2 and the angle of refraction or emergence e. The angle between the emergent ray RS and the direction of the incident ray PQ is called the angle of deviation, δ.

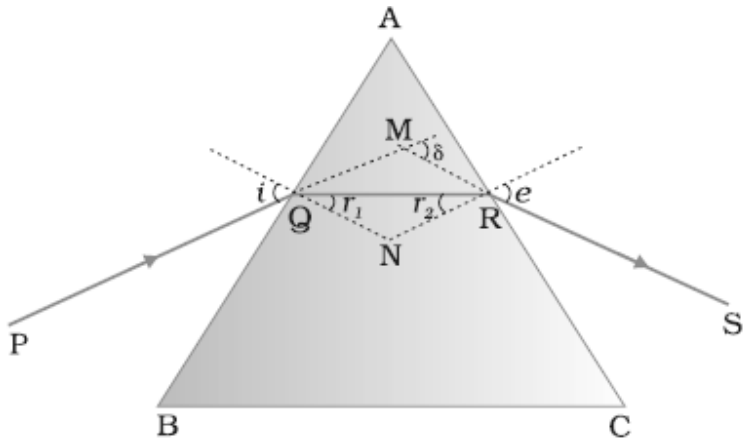


Fig. 2.23 A ray of light passing through a triangular glass prism.

In the quadrilateral AQNR, two of the angles (at the vertices Q and R) are right angles. Therefore, the sum of the other angles of the quadrilateral is 180°.

$\angle A + \angle QNR = 180^\circ$

From the triangle QNR,

$$r_1 + r_2 + \angle QNR = 180^\circ$$

Comparing these two equations, we get

$$r_1 + r_2 = A \tag{2.34}$$

The total deviation δ is the sum of deviation sat the two faces,

$$\delta = (i - r_1) + (e - r_2)$$

that is,

$$\delta = i + e - A \tag{2.35}$$

Thus, the angle of deviation depends on the angle of incidence. A plot between the angle of deviation and angle of incidence is shown in Fig. 2.23. You can see that, in general, any given value of δ, except for i = e, corresponds to two values i and hence of e. This, in fact, is expected from the symmetry of i and e in Eq. (2.35), i.e., δ remains the same if i and e are interchanged. Physically, this is related to the fact that the path of ray in Fig. 2.22 can be traced back, resulting in the same angle of deviation. At the minimum deviation D_m, the refracted ray inside the prism becomes parallel to its base. We have

$\delta = D_m$, i = e which implies $r_1 = r_2$.

Equation (2.34) gives

$$2_r = A \; or \; r = A/2 \tag{2.36}$$

In the same way, Eq. (2.35) gives

$D_m = 2i - A$, or $i = (A + D_m)/2$ (2.37)

The refractive index of the prism is

$$n_{21} = \frac{n_2}{n_1} = \frac{\sin[(A+D_m)/2]}{\sin[A/2]} \tag{2.38}$$

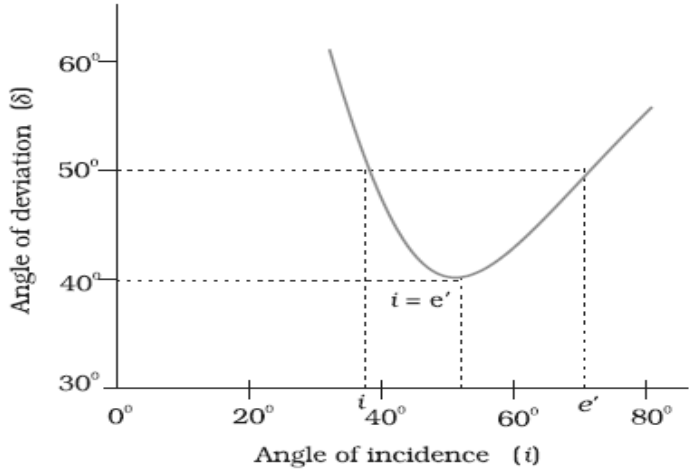


Fig. 2.24 Plot of angle of deviation (δ) versus angle of incidence (*i*) for a triangular prism.

The angles A and D_m can be measured experimentally. Equation (2.38) thus provides a method of determining refractive index of the material of the prism. For a small angle prism, i.e., a thin prism, D_m is also very small, and we get

$$n_{21} = \frac{\sin[(A + D_m)/2]}{\sin[A/2]} \cong \frac{(A + D_m)/2}{A/2}$$

$$D_m = (n_{21} - 1)A$$

It implies that, thin prisms do not deviate light much.

2.7 SOME NATURAL PHENOMENA DUE TO SUNLIGHT

The interplay of light with things around us gives rise to several beautiful phenomena. The spectacle of colour that we see around us all the time is possible only due to sunlight.

While studying dispersion of visible (or white) light by a prism and the electromagnetic spectrum we got to know that colour is associated with frequency of or the wavelength of light in the given medium. In the visible spectrum, red light is at the long wavelength end (~700 nm) while the violet light is at the short wavelength end (~ 400 nm). Dispersion takes place because the refractive index of medium for different wavelengths (colours) is different. For example, the bending of red component of white light is least while it is most for the violet. Equivalently, red light travels faster than violet light in a glass prism. Table 2.2 gives the refractive indices for different wavelength for crown glass and flint glass. Thick lenses could be assumed as made of many prisms, therefore, thick lenses show *chromatic aberration* due to dispersion of light. When white light passes through thick lenses, red and blue colours focus at different points. This phenomenon is known as *chromatic aberration*

TABLE 2.2 REFRACTIVE INDICES FOR DIFFERENT WAVELENGTHS

Colour	Wavelength (nm)	Crown glass	Flint glass
Violet	396.9	1.533	1.663
Blue	486.1	1.523	1.639
Yellow	589.3	1.517	1.627
Red	656.3	1.515	1.622

The variation of refractive index with wavelength may be more pronounced in some media than the other. In vacuum, of course, the speed of light is independent of wavelength. Thus, vacuum (or air approximately) is a non-dispersive medium in which all colours travel with the same speed. This also follows from the fact that sunlight reaches us in the form of white light and not as its components. On the other hand, glass is a dispersive medium.

The blue of the sky, white clouds, the red-hue at sunrise and sunset, the rainbow, the brilliant colours of some pearls, shells, and wings of birds, are just a few of the natural wonders we are used to. We describe some of them here from the point of view of physics.

2.7.1 The rainbow

The rainbow is an example of the dispersion of sunlight by the water drops in the atmosphere. This is a phenomenon due to combined effect of dispersion, refraction and reflection of sunlight by spherical water droplets of rain. The conditions for observing a rainbow are that the sun should be shining in one part of the sky (say near western horizon) while it is raining in the opposite part of the sky (say eastern horizon). An observer can therefore see a rainbow only when his back is towards the sun.

In order to understand the formation of rainbows, consider Fig. (2.24(a). Sunlight is first refracted as it enters a raindrop, which causes the different wavelengths (colours) of white light to separate. Longer wavelength of light (red) are bent the least while the shorter wavelength (violet) are bent the most. Next, these component rays strike the inner surface of the water drop and get internally reflected if the angle between the refracted ray and normal to the drop surface is greater then the critical angle (48º, in this case). The reflected light is refracted again as it comes out of the drop as shown in the figure. It is found that the violet light emerges at an angle of 40º related to the incoming sunlight and red light emerges at an angle of 42º. For other colours, angles lie in between these two values.

Figure 2.24 explains the formation of primary rainbow. We see that red light from drop 1 and violet light from drop 2 reach the observers eye. The violet from drop 1 and red light from drop 2 are directed at level above or below the observer. Thus the observer sees a rainbow with red colour on the top and violet on the bottom. Thus, the primary rainbow is a result of three-step process, that is, refraction, reflection and refraction.

When light rays undergoes *two* internal reflections inside a raindrop, instead of one as in the primary rainbow, a secondary rainbow is formed as shown in Fig. 2.24(c). It is due to four-step process. The intensity of light is reduced at the second reflection and hence the secondary rainbow is fainter than the primary rainbow. Further, the order of the colours is reversed in it as is clear from Fig. 2.24(c).

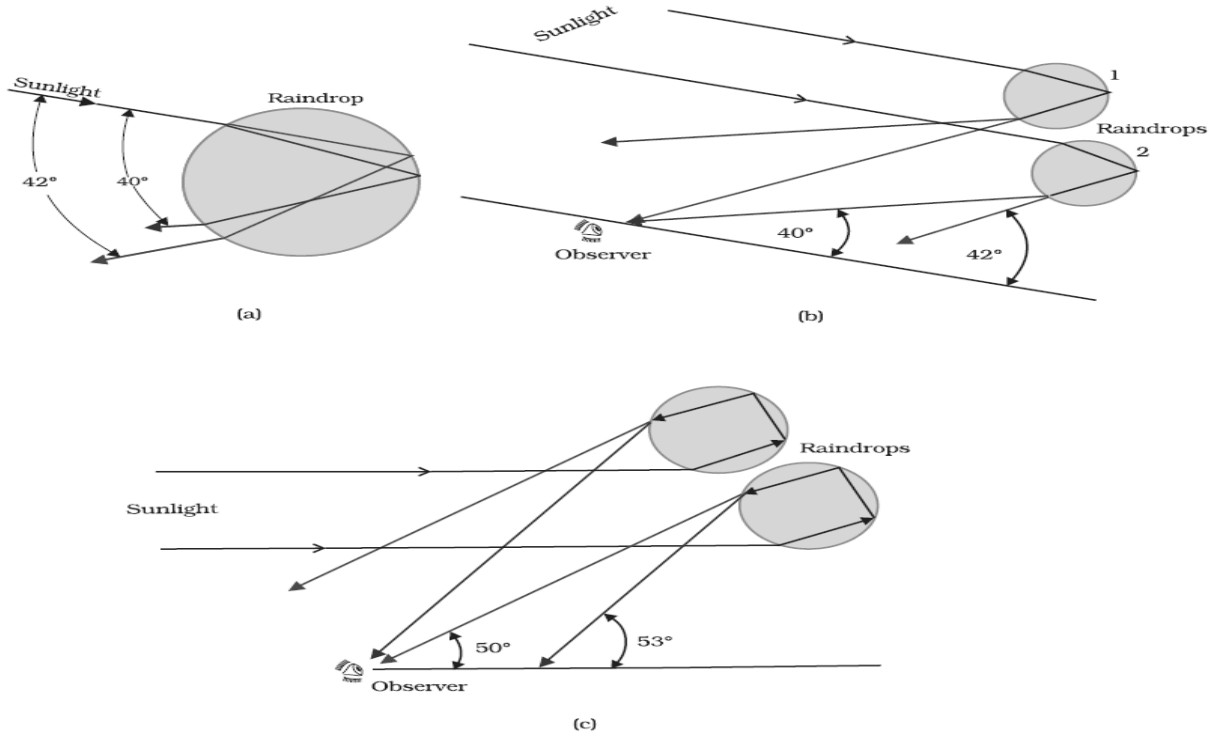


Fig. 2.27 Rainbow: (a) The sun rays incident on a water drop get refracted twice and reflected internally by a drop; (b) Enlarge view of internal reflection and refraction of a ray of light inside a drop form primary rainbow; and (c) secondary rainbow is formed by rays undergoing internal reflection twice inside the drop.

2.7.2 Scattering of light

As sunlight travels through the earth's atmosphere, it gets scattered (changes its direction) by the atmospheric particles. Light of shorter wavelengths is scattered much more than light of longer wavelengths. (The amount of scattering is inversely proportional to the fourth power of the wavelength. This is known as Rayleigh scattering). Hence, the bluish colour predominates in a clear sky, since blue has a shorter wavelength than red and is scattered much more strongly. In fact, violet gets scattered even more than blue, having a shorter wavelength. But since our eyes are more sensitive to blue than violet, we see the sky blue.

Large particles like dust and water droplets present in the atmosphere behave differently. The relevant quantity here is the relative size of the wavelength of light λ, and the scatterer (of typical size, say, a). For a $<< \lambda$, one has Rayleigh scattering which is proportional to $(1/\lambda)^4$. For a $>> \lambda$, i.e., large scattering objects (for example, raindrops, large dust or ice particles) this is not true; all wavelengths are scattered nearly equally. Thus, clouds which have droplets of water with a $>> \lambda$ are generally white.

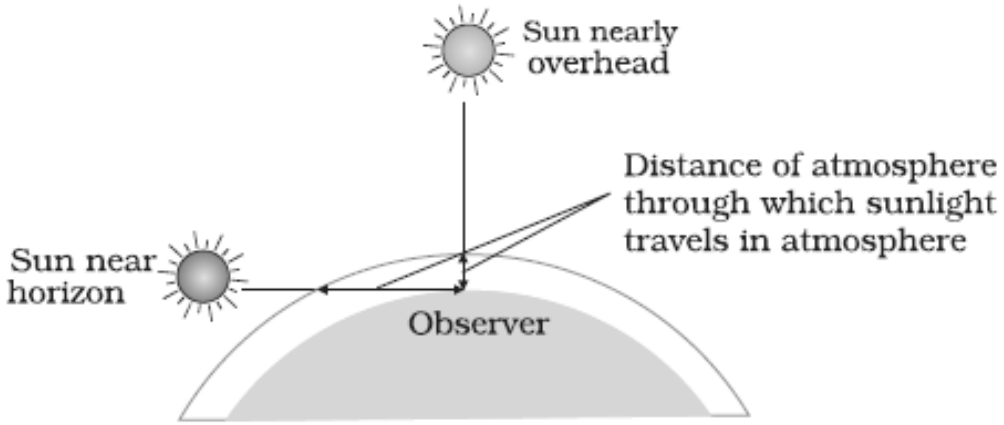


Fig. 2.25 Sunlight travels through a longer distance in the atmosphere at sunset and sunrise.

At sunset or sunrise, the sun's rays have to pass through a larger distance in the atmosphere (Fig. 2.25). Most of the blue and other shorter wavelengths are removed by scattering. The least scattered light reaching our eyes, therefore, the sun looks reddish. This explains the reddish appearance of the sun and full moon near the horizon.

2.8 OPTICAL INSTRUMENTS

A number of optical devices and instruments have been designed utilising reflecting and refracting properties of mirrors, lenses and prisms. Periscope, kaleidoscope, binoculars, telescopes, microscopes are some examples of optical devices and instruments that are in common use. Our eye is, of course, one of the most important optical device the nature has endowed us with. Starting with the eye, we then go on to describe the principles of working of the microscope and the telescope.

2.8.1 The microscope

A simple magnifier or microscope is a converging lens of small focal length (Fig. 2.26). In order to use such a lens as a microscope, the lens is held near the object, one focal length away or less, and the eye is positioned close to the lens on the other side. The idea is to get an erect, magnified and virtual image of the object at a distance so that it can be viewed comfortably, i.e.,

at 25 cm or more. If the object is at a distance f, the image is at infinity. However, if the object is at a distance slightly less than the focal length of the lens, the image is virtual and closer than infinity. Although the closest comfortable distance for viewing the image is when it is at the near point (distance D ≅ 25 cm), it causes some strain on the eye. Therefore, the image formed at infinity is often considered most suitable for viewing by the relaxed eye. We show both cases, the first in Fig. 2.26(a), and the second in Fig. 2.27(b) and (c).

The linear magnification m, for the image formed at the near point D, by a simple microscope can be obtained by using the relation

$$m = \frac{u}{v} = v\left(\frac{1}{v} - \frac{1}{f}\right) = \left(1 - \frac{v}{f}\right)$$

Now according to our sign convention, v is negative, and is equal in magnitude to D. Thus, the magnification is

$$m = \left(1 + \frac{D}{f}\right) \tag{2.39}$$

Since D is about 25 cm, to have a magnification of six, one needs a convex lens of focal length, f = 5 cm.

Note that m = h′/h where h is the size of the object and h′ the size of the image. This is also the ratio of the angle subtended by the image to that subtended by the object, if placed at D for comfortable viewing. (Note that this is not the angle actually subtended by the object at the eye, which is h/u.) What a single-lens simple magnifier achieves is that it allows the object to be brought closer to the eye than D.

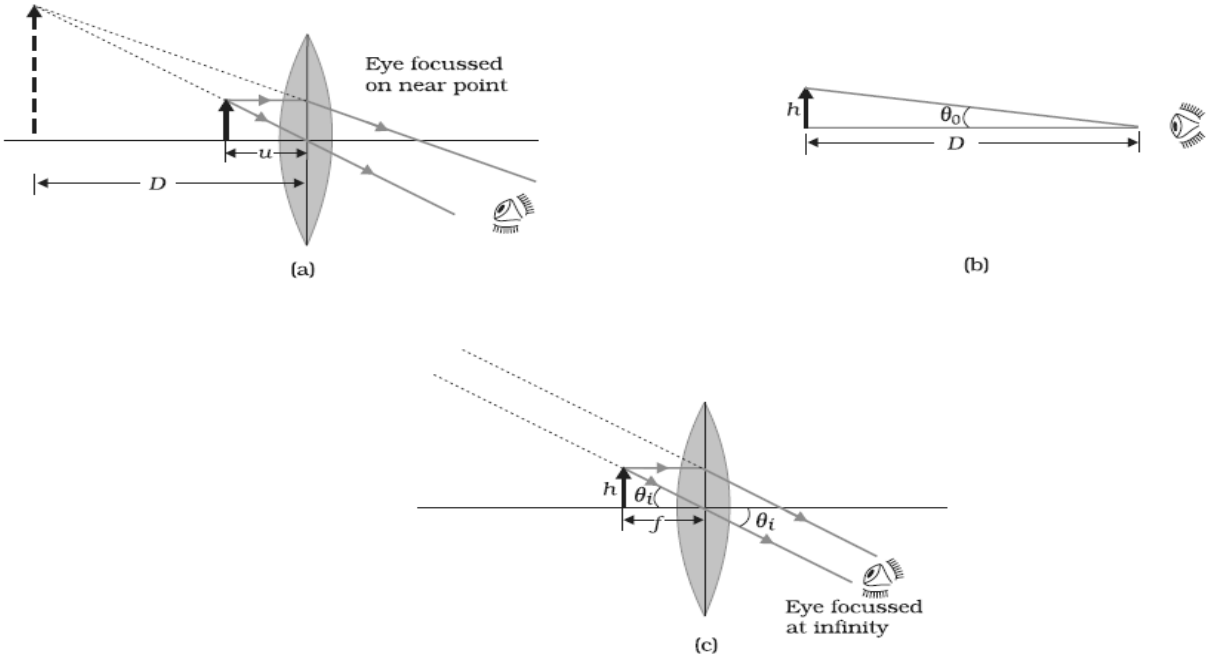


Fig. 2.26 A simple microscope; (a) the magnifying lens is located such that the image is at the near point, (b) the angle subtanded by the object, is the same as that at the near point, and (c) the object near the focal point of the lens; the image is far off but closer than infinity.

We will now find the magnification when the image is at infinity. In this case we will have to obtained the angular magnification. Suppose the object has a height h. The maximum angle it can subtend, and be clearly visible (without a lens), is when it is at the near point, i.e., a distance D. The angle subtended is then given by

$$\tan\theta_o = \left(\frac{}{D}\right) \approx \theta_o \tag{2.40}$$

We now find the angle subtended at the eye by the image when the object is at u. From the relations

$$\frac{|}{} = m = \frac{v}{u}$$

we have the angle subtended by the image

$\tan\theta_i = \frac{|}{-v} = \frac{}{-v}.\frac{v}{u} = \frac{}{-u} \approx \theta$. The angle subtended by the object, when it is at $u = -f$

$$\theta_i = \left(\frac{}{f}\right) \tag{2.41}$$

as is clear from Fig. 2.26(c). The angular magnification is, therefore

$$m = \left(\frac{\theta_i}{\theta_o}\right) = \frac{D}{f} \tag{2.42}$$

This is one less than the magnification when the image is at the near point, Eq. (2.39), but the viewing is more comfortable and the difference in magnification is usually small. In subsequent discussions of optical instruments (microscope and telescope) we shall assume the image to be at infinity.

A simple microscope has a limited maximum magnification (≤ 9) for realistic focal lengths. For much larger magnifications, one uses two lenses, one compounding the effect of the other. This is known as a compound microscope. A schematic diagram of a compound microscope is shown in Fig. 2.27. The lens nearest the object, called the objective, forms a real, inverted, magnified image of the object. This serves as the object for the second lens, the eyepiece, which functions essentially like a simple microscope or magnifier, produces the final image, which is enlarged and virtual. The first inverted image is thus near (at or within) the focal plane of the eyepiece, at a distance appropriate for final image formation at infinity, or a little closer for image formation at the near point. Clearly, the final image is inverted with respect to the original object.

We now obtain the magnification due to a compound microscope. The ray diagram of Fig. 2.27 shows that the (linear) magnification due to the objective, namely h`/h, equals

$$m_o = \frac{|}{} = \frac{L}{f_o}$$

Where we have used the result

$$\tan\beta = \left(\frac{}{f_o}\right) = \left(\frac{|}{L}\right)$$

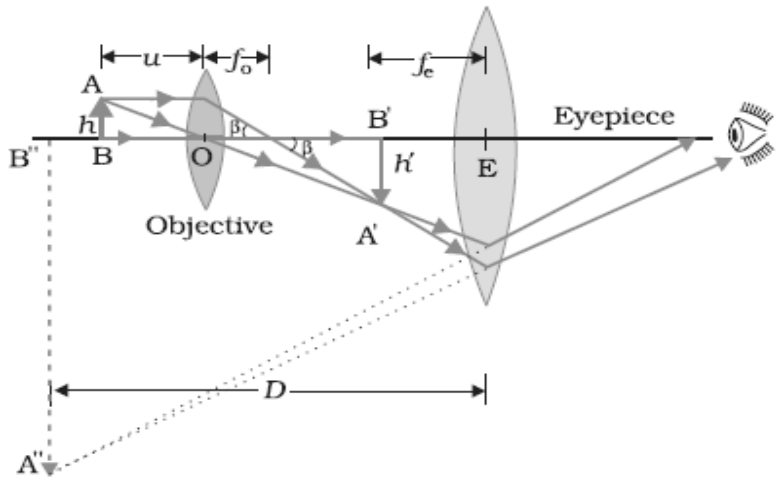


Fig. 2.27 Ray diagram for the formation of image by a compound compound microscope.

Here h′ is the size of the first image, the object size being h and f_o being the focal length of the objective. The first image is formed near the focal point of the eyepiece. The distance L, i.e., the distance between the second focal point of the objective and the first focal point of the eyepiece (focal length f_e) is called the tube length of the compound microscope.

As the first inverted image is near the focal point of the eyepiece, we use the result from the discussion above for the simple microscope to obtain the (angular) magnification me due to it [Eq. (2.39)], when the final image is formed at the near point, is

$$m_e = \left(1 + \frac{D}{f_e}\right) \qquad [2.44\ (a)]$$

When the final image is formed at infinity, the angular magnification due to the eyepiece [Eq. (2.42)] is

$$m_e = (D/f_e) \qquad [2.44(b)]$$

Thus, the total magnification [(according to Eq. (2.33)], when the image is formed at infinity, is

$$m = m_o m_e = \left(\frac{L}{f_o}\right)\left(\frac{D}{f_e}\right) \qquad (2.45)$$

Clearly, to achieve a large magnification of a small object (hence the name microscope), the objective and eyepiece should have small focal lengths. In practice, it is difficult to make the focal length much smaller than 1 cm. Also large lenses are required to make L large. For example, with an objective with fo = 1.0 cm, and an eyepiece with focal length f_e = 2.0 cm, and a tube length of 20 cm, the magnification is

$$m = m_o m_e = \left(\frac{L}{f_o}\right)\left(\frac{D}{f_e}\right)$$

$$= \frac{20}{1} \times \frac{25}{2} = 250$$

Various other factors such as illumination of the object, contribute to the quality and visibility of the image. In modern microscopes, multi-component lenses are used for both the objective and the eyepiece to improve image quality by minimising various optical aberrations (defects) in lenses.

2.8.2 Telescope

The telescope is used to provide angular magnification of distant objects (Fig. 2.28). It also has an objective and an eyepiece. But here, the objective has a large focal length and a much larger aperture than the eyepiece. Light from a distant object enters the objective and a real image is formed in the tube at its second focal point. The eyepiece magnifies this image producing a final inverted image. The magnifying power m is the ratio of the angle β subtended at the eye by the final image to the angle α which the object subtends at the lens or the eye. Hence

$$m \approx \frac{\beta}{\alpha} \approx \frac{\square}{f_e} \cdot \frac{f_o}{\square} = \frac{f_o}{f_e} \tag{2.46}$$

In this case, the length of the telescope tube is $f_o + f_e$.

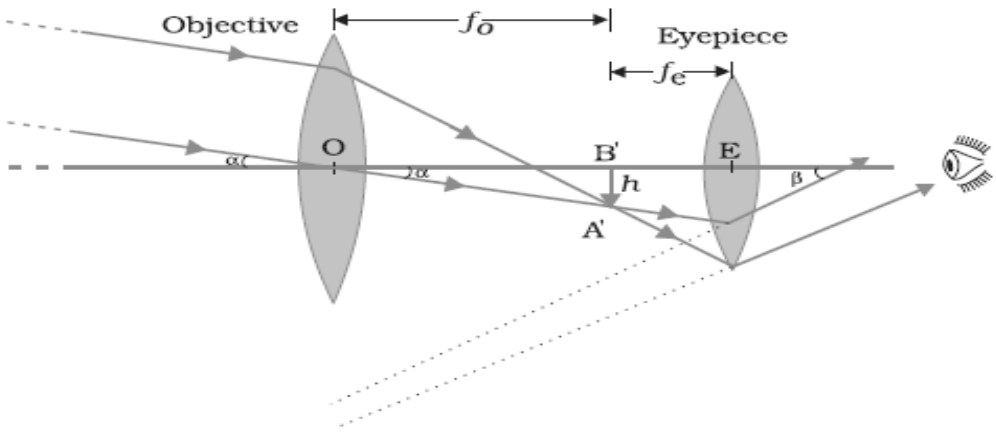


Fig. 2.28 A refracting telescope.

Terrestrial telescopes have, in addition, a pair of inverting lenses to make the final image erect. Refracting telescopes can be used both for terrestrial and astronomical observations. For example, consider a telescope whose objective has a focal length of 100 cm and the eyepiece a focal length of 1 cm. The magnifying power of this telescope is m = 100/1 = 100.

Let us consider a pair of stars of actual separation 1′ (one minute of arc). The stars appear as though they are separated by an angle of 100 × 1′ = 100′ =1.67º.

The main considerations with an astronomical telescope are its light gathering power and its resolution or resolving power. The former clearly depends on the area of the objective. With larger diameters, fainter objects can be observed. The resolving power, or the ability to observe two objects distinctly, which are in very nearly the same direction, also depends on the diameter of the objective. So, the desirable aim in optical telescopes is to make them with objective of large diameter. The largest lens objective in use has a diameter of 40 inch (~1.02 m). It is at the Yerkes Observatory in Wisconsin, USA. Such big lenses tend to be very heavy and therefore, difficult to make and support by their edges. Further, it is rather difficult and expensive to make such large sized lenses which form images that are free from any kind of chromatic aberration and distortions.

For these reasons, modern telescopes use a concave mirror rather than a lens for the objective. Telescopes with mirror objectives are called reflecting telescopes. They have several advantages. First, there is no chromatic aberration in a mirror. Second, if a parabolic reflecting surface is chosen, spherical aberration is also removed. Mechanical support is much less of a problem since a mirror weighs much less than a lens of equivalent optical quality, and can be supported over its entire back surface, not just over its rim. One obvious problem with a reflecting telescope is that the objective mirror focusses light inside the telescope tube. One must have an eyepiece and the observer right there, obstructing some light (depending on the size of the observer cage). This is what is done in the very large 200 inch (~5.08 m) diameters, Mt. Palomar telescope, California. The viewer sits near the focal point of the mirror, in a small cage. Another solution to the problem is to deflect the light being focussed by another mirror. One such arrangement using a convex secondary mirror to focus the incident light, which now passes through a hole in the objective primary mirror, is shown in Fig. 2.29.

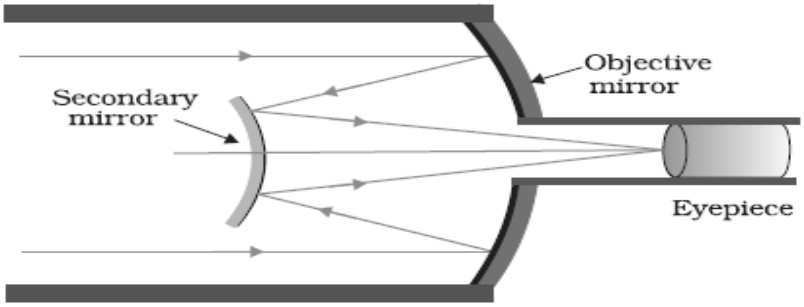


Fig. 2.29 Schematic diagram of a reflecting telescope (Cassegrain).

This is known as a Cassegrain telescope, after its inventor. It has the advantages of a large focal length in a short telescope. The largest telescope in India is in Kavalur, Tamil Nadu. It is a 2.34 m diameter reflecting telescope (Cassegrain). It was ground, polished, set up, and is being used by the Indian Institute of Astrophysics, Bangalore. The largest reflecting telescopes in the world are the pair of Keck telescopes in Hawaii, USA, with a reflector of 10 metre in diameter.

SUMMARY

1. Reflection is governed by the equation $\angle i = \angle r'$ and refraction by the Snell's law, $\sin i/\sin r = n$, where the incident ray, reflected ray, refracted ray and normal lie in the same plane. Angles of incidence, reflection and refraction are i, r ′ and r, respectively.
2. The critical angle of incidence ic for a ray incident from a denser to rarer medium, is that angle for which the angle of refraction is 90°. For $i > i_c$, total internal reflection occurs. Multiple internal reflections in diamond ($i_c \cong 24.4°$), totally reflecting prisms and mirage, are someexamples of total internal reflection. Optical fibres consist of glassfibres coated with a thin layer of material of lower refractive index. Light incident at an angle at one end comes out at the other, after multiple internal reflections, even if the fibre is bent.
3. Cartesian sign convention: Distances measured in the same direction as the incident light are positive; those measured in the opposite direction are negative. All distances are measured from the pole/optic centre of the mirror/lens on the principal axis. The heights measure upwards above x-axis and normal to the principal axis of the mirror/lens are taken as positive. The heights measured downwards are taken as negative.
4. Mirror equation:

$$\frac{1}{v}+\frac{1}{u}=\frac{1}{f}$$

where u and v are object and image distances, respectively and f is the focal length of the mirror. f is (approximately) half the radius of curvature R. f is negative for concave mirror; f is positive for a convex mirror.

5. For a prism of the angle A, of refractive index n_2 placed in a medium of refractive index n_1,

$$n_{21}=\frac{n_2}{n_1}=\frac{\sin[(A+D_m)/2]}{sin(A/2)}$$

where D_m is the angle of minimum deviation.

6. For refraction through a spherical interface (from medium 1 to 2 of refractive index n_1 and n_2, respectively)

$$\frac{n_2}{v}-\frac{n_1}{u}=\frac{n_2-n_1}{R}$$

Thin lens formula

$$\frac{1}{v}-\frac{1}{u}=\frac{1}{f}$$

Lens maker's fomula

$$\frac{1}{f}=\frac{n_2-n_1}{n_1}\left[\frac{1}{R_1}-\frac{1}{R_2}\right]$$

R_1 and R_2 are the radii of curvature of the lens surfaces. f is positive for a converging lens; f is negative for diverging lens. The power of lens P=1/f.

The SI unit for power of a lens is diopter (D): 1 D =1 m^{-1}

If several thin lenses of focal length f1,f2,f3,… are in contact, the effective focal length of their combination is given by

$$\frac{1}{f}=\frac{1}{f_1}+\frac{1}{f_2}+\frac{1}{f_3}+\cdots$$

The total power of a combination of several lenses is

$P=P_1+P_2+P_3+\ldots.$

7. Dispersion is the splitting of light into its constituent colour.

8. Magnifying power m of a simple microscope is given by m=1+(D/f), where D=25cm is the least distance of distinct vision and f is the focal length of the convex lens. If the image is at infinity, $m = D/f$. For a compound microscope, the magnifying power is given by $m = m_e \times m_0$ where $m_e = 1 + (D/f_e)$, is the magnification due to the eyepiece and m_o is the magnification produced by the objective.

Approximately, $$m = \frac{L}{f_0} X \frac{D}{f_e}$$

where f_o and f_e are the focal lengths of the objective and eyepiece, respectively, and L is the distance between their focal points.

9. Magnifying power m of a telescope is the ratio of the angle β subtended at the eye by the image to the angle α subtended at the eye by the object.

$$m=\frac{\beta}{\alpha}=\frac{f_0}{f_e}$$

where f_0 and f_e are the focal lengths of the objective and eyepiece, respectively.

VERY SHORT ANSWER QUESTIONS (2 MARKS)

1. Define focal length and radius of curvature of a concave lens.
2. What do you understand by the terms 'focus' and 'principal focus' in the context of lenses ?
3. What are the laws of reflection through curved mirrors ?
4. Define 'power' of a convex lens. What is its unit ?
5. A small angled prism of 4^0 deviates a ray through 2.48^0 . Find the refractive index of the prism.
6. What is 'dispersion'? Which colour gets relatively more dispersed ?
7. What is myopia ? How can it be corrected ?
8. What is hypermetropia ?How can it be corrected ?

SHORT ANSWER QUESTIONS (4 MARKS)

1. Define focal length of a concave mirror, Prove that the radius of curvature of a concave mirror is double its focal length.
2. Explain the Cartesian sign convention for mirrors.
3. Define critical angle. Explain total internal reflection using a neat diagram.
4. Explain the formation of a mirage
5. Explain the formation of a rainbow.
6. Why does the setting sun appear red ?
7. With a neat labelled diagram explain the formation of image in a simple microscope.
8. What is the position of the object for a simple microscope? What is the maximum magnification of a simple microscope for a realistic focal length ?

LONG ANSWER QUESTIONS (8 MARKS)

1.(a) What is the Cartesian sign convention ? Applying this convention and using a neat diagram, derive an expression for finding the image distance using the mirror equation.

(b) An object of 5cm height is placed at a distance of 15 cm from a concave mirror of radius of curvature 20 cm. Find the size of the image.

2. (a) Using a neat labelled diagram derive the mirror equation. Define linear magnification.

(b) An object is placed at 5 cm from a convex lens of focal length 15cm. What is the position and nature of the image?

3. (a) Define Snell's Law. Using a neat labelled diagram derive an expression for the refractive index of the material of an equilateral prism.

(b) A ray of light, after passing through a medium, meets the surface separating the medium from air at an angle of 45^0 and is just not reflected. What is the refractive index of the medium ?

4. Draw a neat labelled diagram of a compound microscope and explain its working. Derive an expression for its magnification.

CHAPTER 3

WAVE OPTICS

3.1 INTRODUCTION

In 1637 Descartes gave the corpuscular model of light and derived Snell's law. It explained the laws of reflection and refraction of light at an interface. The corpuscular model predicted that if the ray of light (on refraction) bends towards the normal then the speed of light would be greater in the second medium. This corpuscular model of light was further developed by Isaac Newton in his famous book entitled *OPTICKS* and because of the tremendous popularity of this book, the corpuscular model is very often attributed to Newton. In 1678, the Dutch physicist Christiaan Huygens put forward the wave theory of light – it is this wave model of light that we will discuss in this chapter. As we will see, the wave model could satisfactorily explain the phenomena of reflection and refraction; however, it predicted that on refraction if the wave bends towards the normal then the speed of light would be less in the second medium. This is in contradiction to the prediction made by using the corpuscular model of light. It was much later confirmed by experiments where it was shown that the speed of light in water is less than the speed in air confirming the prediction of the wave model; Foucault carried out this experiment in 1850.

The wave theory was not readily accepted primarily because of Newton's authority and also because light could travel through vacuum and it was felt that a wave would always requires a medium to propagate from one point to the other. However, when Thomas Young performed his famous interference experiment in 1801, it was firmly established that light is indeed a wave phenomenon. The wavelength of visible light was measured and found to be extremely small; for example, the wavelength of yellow light is about 0.5 μm. Because of the smallness of the wavelength of visible light (in comparison to the dimensions of typical mirrors and lenses), light can be assumed to approximately travel in straight lines. This is the field of geometrical optics, which we had discussed in the previous chapter. Indeed, the branch of optics in which one completely neglects the finiteness of the wavelength is called geometrical optics and a ray is defined as the path of energy propagation in the limit of wavelength tending to zero.

After the interference experiment of Young in 1801, for the next 40 years or so, many experiments were carried out involving the interference and diffraction of light waves; these experiments could only be satisfactorily explained by assuming a wave model of light. Thus, around the middle of the nineteenth century, the wave theory seemed to be very well established. The only major difficulty was that since it was thought that a wave required a medium for its propagation, how light waves could propagate through vacuum. This was explained when Maxwell put forward his famous electromagnetic theory of light. Maxwell had developed a set of equations describing the laws of electricity and magnetism and using these equations he derived what is known as the wave equation from which he predicted the existence of electromagnetic waves*. From the wave equation, Maxwell could calculate the speed of electromagnetic waves in free space and he found that the theoretical value was very close to the measured value of speed of light. From this, he propounded that light must be an electromagnetic wave. Thus, according to Maxwell, light waves are associated with changing electric and magnetic fields; changing electric field produces a time and space varying magnetic field and a changing magnetic field produces a time and space varying electric field. The changing electric and

magnetic fields result in the propagation of electromagnetic waves (or light waves) even in vacuum.

In this chapter we will first discuss the original formulation of the Huygens principle and derive the laws of reflection and refraction. In Sections 3.4 and 3.5, we will discuss the phenomenon of interference which is based on the principle of superposition. In Section 3.6 we will discuss the phenomenon of diffraction which is based on Huygens-Fresnel principle. Finally in Section 3.7 we will discuss the phenomenon of polarisation which is based on the fact that the light waves are *transverse electromagnetic waves*.

3.2 HUYGENS PRINCIPLE

We would first define a wavefront: when we drop a small stone on a calm pool of water, waves spread out from the point of impact. Every point on the surface starts oscillating with time. At any instant, a photograph of the surface would show circular rings on which the disturbance is maximum. Clearly, all points on such a circle are oscillating in phase because they are at the same distance from the source. Such a locus of points, which oscillate in phase is called a wavefront; thus a wavefront is defined as a surface of constant phase.

The speed with which the wavefront moves outwards from the source is called the speed of the wave. The energy of the wave travels in a direction perpendicular to the wave front.

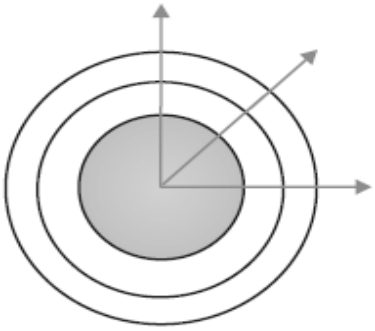


Fig. 3.1 (a) A diverging spherical wave emanating from a point source. The wave fronts are spherical.

If we have a point source emitting waves uniformly in all directions, then the locus of points which have the same amplitude and vibrate in the same phase are spheres and we have what is known as a spherical wave as shown in Fig. 3.1(a). At a large distance from the source, a small portion of the sphere can be considered as a plane and we have what is known as a plane wave [Fig. 3.1(b)].

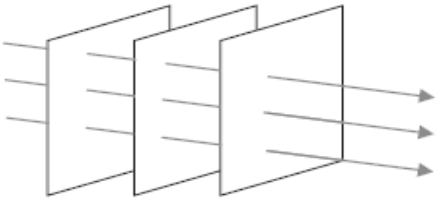


Fig. 3.1(b) At the large distance from the source, a small portion of the spherical wave can be approximated by a plane wave.

---

* Maxwell had predicted the existence of electromagnetic waves around 1855; it was much later around (1890) that Heinrich Hertz produced radio waves in the laboratory. J.C Bose and G. Marconi made particular applications of the *Hertzian waves.*

Now, if we know the shape of the wavefront at t = 0, then Huygens principle allows us to determine the shape of the wavefront at a later time τ. Thus, Huygens principle is essentially a geometrical construction, which given the shape of the wafefront at any time allows us to determine the shape of the wavefront at a later time. Let us consider a diverging wave and let F_1F_2 represent a portion of the spherical wavefront at t = 0 (Fig. 3.2). Now, according to Huygens principle, each point of the wavefront is the source of a secondary disturbance and the wavelets emanating from these points spread out in all directions with the speed of the wave. These wavelets emanating from the wavefront are usually referred to as secondary wavelets and if we draw a common tangent to all these spheres, we obtain the new position of the wavefront at a later time.

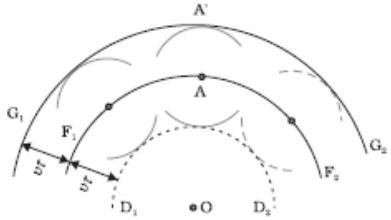


Fig. 3.2 F_1F_2 represents the spherical wavefront (with O as centre) at t=0. The envelope of the secondary wavelets emanating from F_1F_2 produces the forward moving wavefront G_1G_2. The backwave D_1D_2 does not exist.

Thus, if we wish to determine the shape of the wavefront at t = τ, we draw spheres of radius vτ from each point on the spherical wavefront where v represents the speed of the waves in the medium. If we now draw a common tangent to all these spheres, we obtain the new position of the wavefront at t = τ. The new wavefront shown as G_1G_2 in Fig. 3.2 is again spherical with point O as the centre.

The above model has one shortcoming: we also have a back wave which is shown as D_1D_2 in Fig. 3.2. Huygens argued that the amplitude of the secondary wavelets is maximum in the forward direction and zero in the backward direction; by making this adhoc assumption, Huygens could explain the absence of the backwave. However, this adhoc assumption is not satisfactory and the absence of the backwave is really justified from more rigorous wave theory.

In a similar manner, we can use Huygens principle to determine the shape of the wavefront for a plane wave propagating through a medium (Fig. 3.3).

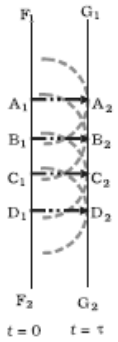


Fig. 3.3 Huygens geometrical construction for a plane wave propagating to the right. F_1F_2 is the plane wavefront at t=0 and G_1G_2 is the wavefront at a later time τ. The line A_1A_2, B_1B_2……etc. are normal to both F_1F_2 and G_1G_2 and represent rays.

3.3 REFRACTION AND REFLECTION OF PLANE WAVES USING HUYGENS PRINCIPLE

3.3.1 Refraction of a plane wave

We will now use Huygens principle to derive the laws of refraction. Let PP' represent the surface separating medium 1 and medium 2, as shown in Fig. 3.4. Let ν_1 and ν_2 represent the speed of light in medium 1 and medium 2, respectively. We assume a plane wavefront AB propagating in the direction A′A incident on the interface at an angle i as shown in the figure. Let τ be the time taken by the wavefront to travel the distance BC. Thus,

$$BC = \nu_1\tau$$

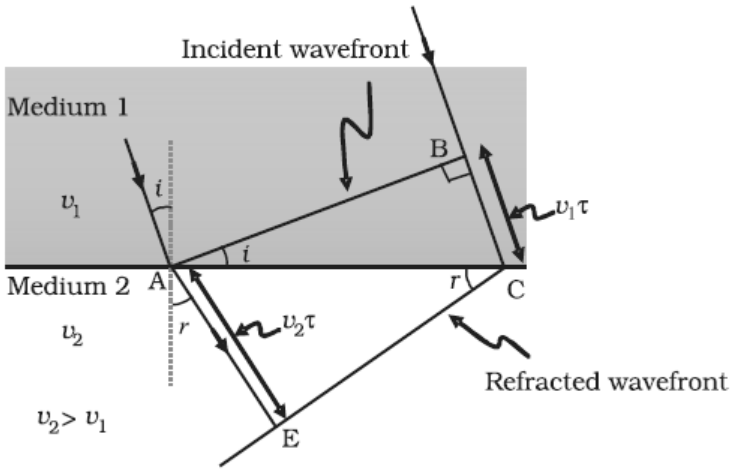


Fig. 3.4 A plane wave AB is incident at an angle i on the surface pp' separating medium 1 and medium 2. The plane wave undergoes refraction and CE represents the refracted wavefront. The figure corresponds to $\vartheta_2 < \vartheta_1$ so that the refracted waves bends towards the normal.

In order to determine the shape of the refracted wavefront, we draw a sphere of radius $\nu_2\tau$ from the point A in the second medium (the speed of the wave in the second medium is ν_2). Let CE represent a tangent plane drawn from the point C on to the sphere. Then, AE = $\nu_2\tau$ and CE would represent the refracted wavefront. If we now consider the triangles ABC and AEC, we readily obtain

$$\sin i = \frac{BC}{AC} = \frac{\nu_1\tau}{AC} \qquad (3.1)$$

and

$$\sin r = \frac{AE}{AC} = \frac{\nu_2\tau}{AC} \qquad (3.2)$$

Where i and r are the angles of incidence and refraction, respectively.

Thus we obtain

$$\frac{\sin i}{\sin r} = \frac{\nu_1}{\nu_2} \qquad (3.3)$$

From the above equation, we get the important result that if r < i (i.e., if the ray bends toward the normal), the speed of the light wave in the second medium (ν_2) will be less than the

speed of the light wave in the first medium (ν_1). This prediction is opposite to the prediction from the corpuscular model of light and as later experiments showed, the prediction of the wave theory is correct. Now, if c represents the speed of light in vacuum, then,

$$n_1 = \frac{c}{\nu_1} \qquad (3.4)$$

and

$$n_2 = \frac{c}{\nu_2} \qquad (3.5)$$

are known as the refractive indices of medium 1 and medium 2, respectively. In terms of the refractive indices, Eq. (3.3) can be written as

$$n_1 \sin i = n_2 \sin r \qquad (3.6)$$

This is the Snell's law of refraction. Further, if λ_1 and λ_2 denote the wavelengths of light in medium 1 and medium 2, respectively and if the distance BC is equal to λ_1 then the distance AE will be equal to λ_2 (because if the crest from B has reached C in time τ, then the crest from A should have also reached E in time τ); thus,

$$\frac{\lambda_1}{\lambda_2} = \frac{BC}{AE} = \frac{\vartheta_1}{\vartheta_2}$$

Or

$$\frac{\vartheta_1}{\lambda_1} = \frac{\vartheta_2}{\lambda_2} \qquad (3.7)$$

The above equation implies that when a wave gets refracted into a denser medium ($\nu_1 > \nu_2$) the wavelength and the speed of propagation decrease but the frequency ν ($=\nu/\lambda$) remains the same.

3.3.2 Refraction at a rarer medium

We now consider refraction of a plane wave at a rarer medium, i.e., $\nu_2 > \nu_1$. Proceeding in an exactly similar manner we can construct a refracted wavefront as shown in Fig. 3.5. The angle of refraction will now be greater than angle of incidence; however, we will still have $n_1 \sin i = n_2 \sin r$. We define an angle i_c by the following equation

$$\sin i_c = \frac{n_2}{n_1} \qquad (3.8)$$

Thus, if $i = i_c$ then $\sin r = 1$ and $r = 90°$. Obviously, for $i > i_c$, there cannot be any refracted wave. The angle i_c is known as the critical angle and for all angles of incidence greater than the critical angle, we will not have any refracted wave and the wave will undergo what is known as total internal reflection. The phenomenon of total internal reflection and its applications was discussed in Section 2.4.

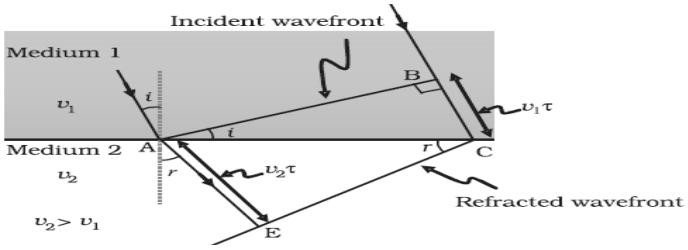


Fig. 3.5 Refraction of a plane wave incident on a rarer medium for which $\vartheta_2 < \vartheta_1$. The plane wave bends away from the normal.

3.3.3 Reflection of a plane wave by a plane surface

We next consider a plane wave AB incident at an angle i on a reflecting surface MN. If ν represents the speed of the wave in the medium and if τ represents the time taken by the wavefront to advance from the point B to C then the distance

$BC = \nu\tau$.

In order the construct the reflected wavefront we draw a sphere of radius $\nu\tau$ from the point A as shown in Fig. 3.6. Let CE represent the tangent plane drawn from the point C to this sphere. Obviously

$AE = BC = \nu\tau$

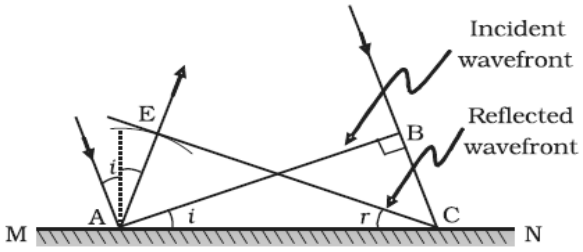


Fig. 3.6 Reflection of a plane wave AB by the reflecting surface MN. AB and CE represent incident and reflected wavefronts.

If we now consider the triangles EAC and BAC we will find that they are congruent and therefore, the angles i and r (as shown in Fig. 3.6) would be equal. This is the law of reflection.

Once we have the laws of reflection and refraction, the behaviour of prisms, lenses, and mirrors can be understood. These phenomena were discussed in detail in Chapter 9 on the basis of rectilinear propagation of light. Here we just describe the behaviour of the wavefronts as they undergo reflection or refraction. In Fig. 3.7(a) we consider a plane wave passing through a thin prism. Clearly, since the speed of light waves is less in glass, the lower portion of the incoming wavefront (which travels through the greatest thickness of glass) will get delayed resulting in a tilt in the emerging wavefront as shown in the figure. In Fig. 3.7(b) we consider a plane wave incident on a thin convex lens; the central part of the incident plane wave traverses the thickest portion of the lens and is delayed the most. The emerging wavefront has a depression at the centre and therefore the wavefront becomes spherical and converges to the point F which is known as the focus. In Fig. 3.7(c) a plane wave is incident on a concave mirror and on reflection we have a spherical wave converging to the focal point F. In a similar manner, we can understand refraction and reflection by concave lenses and convex mirrors.

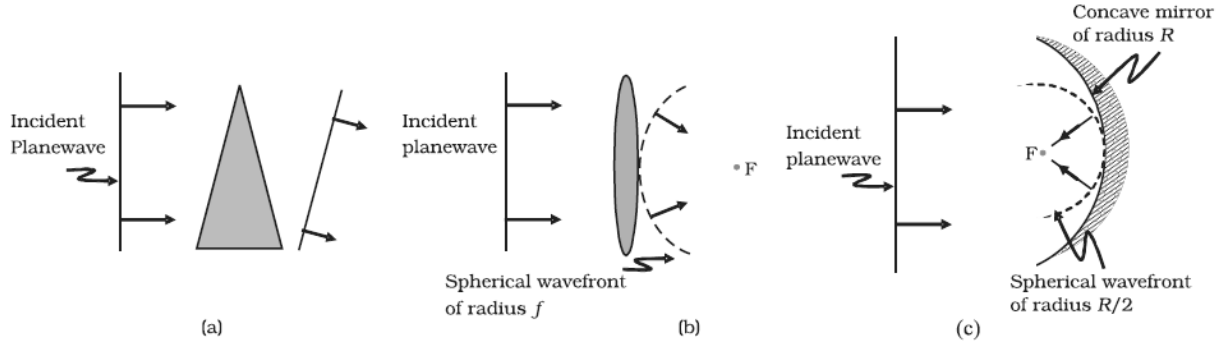


Fig. 3.7 Refraction of a plane wave by (a) a thin prism, (b) a convex lens (c) Reflection of a plane wave by a concave mirror.

From the above discussion it follows that the total time taken from a point on the object to the corresponding point on the image is the same measured along any ray. For example, when a convex lens focuses light to form a real image, although the ray going through the centre traverses a shorter path, but because of the slower speed in glass, the time taken is the same as for rays travelling near the edge of the lens.

3.3.4 The doppler effect

We should mention here that one should be careful in constructing the wavefronts if the source (or the observer) is moving. For example, if there is no medium and the source moves away from the observer, then later wavefronts have to travel a greater distance to reach the observer and hence take a longer time. The time taken between the arrival of two successive wavefronts is hence longer at the observer than it is at the source. Thus, when the source moves away from the observer the frequency as measured by the source will be smaller. This is known as the Doppler effect. Astronomers call the increase in wavelength due to Doppler effect as red shift since a wavelength in the middle of the visible region of the spectrum moves towards the red end of the spectrum. When waves are received from a source moving towards the observer, there is an apparent decrease in wavelength, this is referred to as blue shift.

You have already encountered Doppler effect for sound waves. For velocities small compared to the speed of light, we can use the same formulae which we use for sound waves. The fractional change in frequency $\Delta\nu/\nu$ is given by $-\nu_{radial}/c$, where ν_{radial} is the component of the source velocity along the line joining the observer to the source relative to the observer; ν_{radial} is considered positive when the source moves away from the observer. Thus, the Doppler shift can be expressed as:

$$\frac{\Delta\vartheta}{\vartheta} = -\frac{\vartheta_{radial}}{c} \qquad (3.9)$$

The formula given above is valid only when the speed of the source is small compared to that of light. A more accurate formula for the Doppler effect which is valid even when the speeds are close to that of light, requires the use of Einstein's special theory of relativity. The Doppler effect for light is very important in astronomy. It is the basis for the measurements of the radial velocities of distant galaxies.

3.4 COHERENT AND INCOHERENT ADDITION OF WAVES

In this section we will discuss the interference pattern produced by the superposition of two waves. You may recall that we had discussed the superposition principle. Indeed the entire field of interference is based on the superposition principle according to which at a particular point in the medium, the resultant displacement produced by a number of waves is the vector sum of the displacements produced by each of the waves.

Consider two needles S_1 and S_2 moving periodically up and down in an identical fashion in a trough of water [Fig. 3.8(a)]. They produce two water waves, and at a particular point, the phase difference between the displacements produced by each of the waves does not change with time; when this happens the two sources are said to be coherent. Figure 3.8(b) shows the position of crests (solid circles) and troughs (dashed circles) at a given instant of time. Consider a point P for which

$$S_1P = S_2P$$

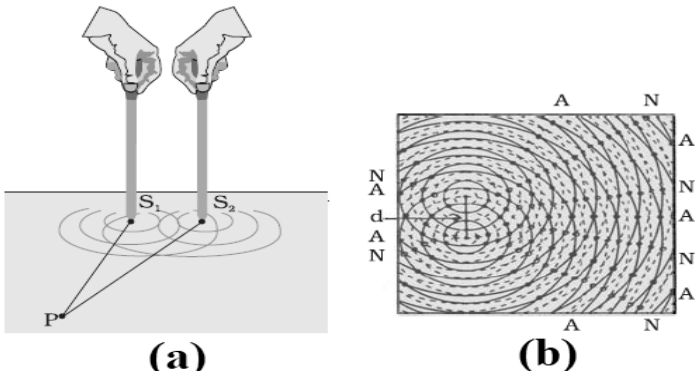


Fig. 3.8 (a) Two needles oscillating in phase in water represent two coherent sources.
(b) The pattern of displacement of water molecules at an instant on the surface of water showing nodal N (no displacement) and antinodal A (maximum displacement) lines.

Since the distances S_1 P and S_2 P are equal, waves from S_1 and S_2 will take the same time to travel to the point P and waves that emanate from S_1 and S_2 in phase will also arrive, at the point P, in phase.

Thus, if the displacement produced by the source S_1 at the point P is given by

$$y_1 = a \cos \omega t$$

then, the displacement produced by the source S2 (at the point P) will also be given by

$$y_2 = a \cos \omega t$$

Thus, the resultant of displacement at P would be given by

$$y = y_1 + y_2 = 2\, a \cos \omega t$$

Since the intensity is the proportional to the square of the amplitude, the resultant intensity will be given by $I = 4\, I_0$

where I_0 represents the intensity produced by each one of the individual sources; I_0 is proportional to a^2. In fact at any point on the perpendicular bisector of S_1S_2, the intensity will be $4I_0$. The two sources are said to interfere constructively and we have what is referred to as constructive interference.

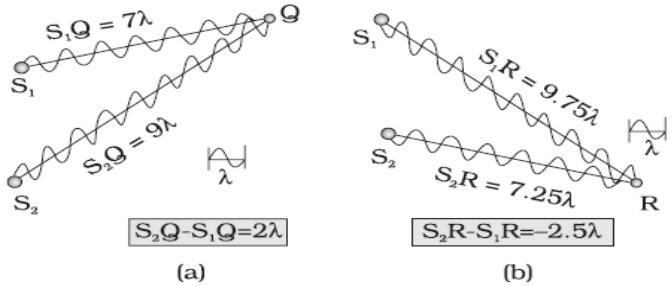


Fig.3.9 (a) Constructive interference at a point Q for which the path difference is 2λ.
(b) Destructive interference at a point R for which the path difference is 2.5 λ.

We next consider a point Q [Fig. 3.9(a)] for which

$$S_2Q - S_1Q = 2\lambda$$

The waves emanating from S1 will arrive exactly two cycles earlier than the waves from S_2 and will again be in phase [Fig. 3.9(a)]. Thus, if the displacement produced by S_1 is given by

$$y_1 = a \cos \omega t$$

then the displacement produced by S_2 will be given by

$$y_2 = a \cos (\omega t - 4\pi) = a \cos \omega t$$

where we have used the fact that a path difference of 2λ corresponds to a phase difference of 4π. The two displacements are once again in phase and the intensity will again be $4I_0$ giving rise to constructive interference. In the above analysis we have assumed that the distances S_1Q and S_2Q are much greater than d (which represents the distance between S_1 and S_2) so that although S_1Q and S_2Q are not equal, the amplitudes of the displacement produced by each wave are very nearly the same.

We next consider a point R [Fig. 3.9(b)] for which

$$S_2R - S_1R = -2.5\lambda$$

The waves emanating from S_1 will arrive exactly two and a half cycles later than the waves from S_2 [Fig. 3.10].

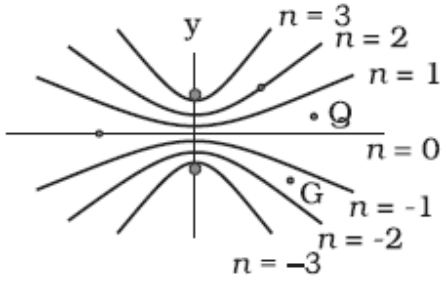


Fig. 3.10 Locus of points for which $S_1P - S_2P$ is equal to zero. $\pm\lambda, \pm 2\lambda, \pm 3\lambda$.

Thus if the displacement produced by S_1 is given by

$$y_1 = a \cos \omega t$$

then the displacement produced by S_2 will be given by

$$y_2 = a \cos (\omega t + 5\pi) = - a \cos \omega t$$

where we have used the fact that a path difference of 2.5λ corresponds to a phase difference of 5π. The two displacements are now out of phase and the two displacements will cancel out to give zero intensity. This is referred to as destructive interference.

To summarise: If we have two coherent sources S_1 and S_2 vibrating in phase, then for an arbitrary point P whenever the path difference,

$$S_1P \sim S_2P = n\lambda \ (n = 0, 1, 2, 3,...) \qquad (3.10)$$

we will have constructive interference and the resultant intensity will be $4I_0$; the sign ~ between S_1P and S_2 P represents the difference between S_1P and S_2P. On the other hand, if the point P is such that the path difference,

$$S_1P \sim S_2P = (n+\tfrac{1}{2})\lambda \ (n = 0, 1, 2, 3, ...) \qquad (3.11)$$

we will have destructive interference and the resultant intensity will be zero. Now, for any other arbitrary point G (Fig. 3.10) let the phase difference between the two displacements be φ. Thus, if the displacement produced by S_1 is given by

$$y_1 = a \cos \omega t$$

then, the displacement produced by S_2 would be

$$y_2 = a \cos (\omega t + \varphi)$$

and the resultant displacement will be given by

$$y = y_1 + y_2$$
$$= a [\cos \omega t + \cos (\omega t + \varphi)]$$
$$= 2 a \cos (\varphi/2) \cos (\omega t + \varphi/2)$$

The amplitude of the resultant displacement is 2a cos (φ/2) and therefore the intensity at that point will be

$$I = 4 I_0 \cos2 (\varphi/2) \qquad (3.12)$$

If φ = 0, ± 2 π, ± 4 π,… which corresponds to the condition given by Eq. (3.10) we will have constructive interference leading to maximum intensity. On the other hand, if φ = ± π, ± 3π, ± 5π … [which corresponds to the condition given by Eq. (3.11)] we will have destructive interference leading to zero intensity.

Now if the two sources are coherent (i.e., if the two needles are going up and down regularly) then the phase difference φ at any point will not change with time and we will have a stable interference pattern; i.e., the positions of maxima and minima will not change with time. However, if the two needles do not maintain a constant phase difference, then the interference pattern will also change with time and, if the phase difference changes very rapidly with time, the positions of maxima and minima will also vary rapidly with time and we will see a "time-averaged" intensity distribution. When this happens, we will observe an average intensity that will be given by

$$< I > = 4I_0 < \cos^2 \varphi/2 > \qquad (3.13)$$

where angular brackets represent time averaging. if φ (t) varies randomly with time, the time-averaged quantity $< \cos^2(\varphi/2) >$ will be 1/2. This is also intuitively obvious because the function $\cos^2(\varphi/2)$ will randomly vary between 0 and 1 and the average value will be 1/2. The resultant intensity will be given by

$$I = 2 I_0 \qquad (3.14)$$

at all points.

When the phase difference between the two vibrating sources changes rapidly with time, we say that the two sources are incoherent and when this happens the intensities just add up. This is indeed what happens when two separate light sources illuminate a wall.

3.5 INTERFERENCE OF LIGHT WAVES AND YOUNG'S EXPERIMENT

We will now discuss interference using light waves. If we use two sodium lamps illuminating two pinholes (Fig. 3.11) we will not observe any interference fringes. This is because of the fact that the light wave emitted from an ordinary source (like a sodium lamp) undergoes abrupt phase changes in times of the order of 10^{-10} seconds. Thus the light waves coming out from two independent sources of light will not have any fixed phase relationship and would be incoherent, when this happens, as discussed in the previous section, the intensities on the screen will add up.

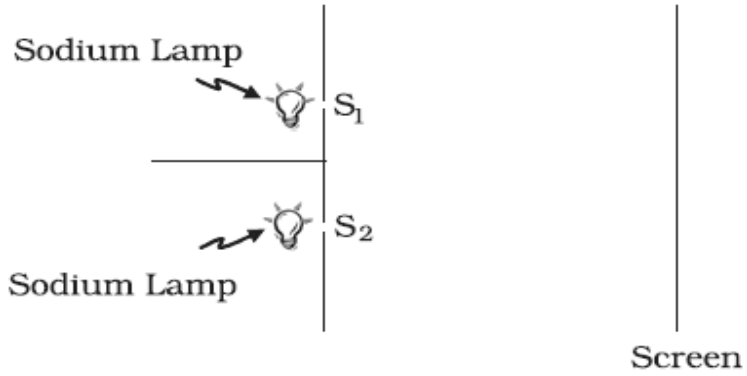


Fig. 3.11 If two sodium lamps illuminate two pinholes S_1 and S_2, the intensities will add up and no interference fringes will be observed on the screen.

The British physicist Thomas Young used an ingenious technique to "lock" the phases of the waves emanating from S_1 and S_2. He made two pinholes S_1 and S_2 (very close to each other) on an opaque screen [Fig. 3.12(a)]. These were illuminated by another pinholes that was in turn, lit by a bright source. Light waves spread out from S and fall on both S_1 and S_2. S_1 and S_2 then behave like two coherent sources because light waves coming out from S_1 and S_2 are derived from the same original source and any abrupt phase change in S will manifest in exactly similar phase changes in the light coming out from S_1 and S_2. Thus, the two sources S_1 and S_2 will be locked in phase; i.e., they will be coherent like the two vibrating needle in our water wave example [Fig. 3.8(a)].

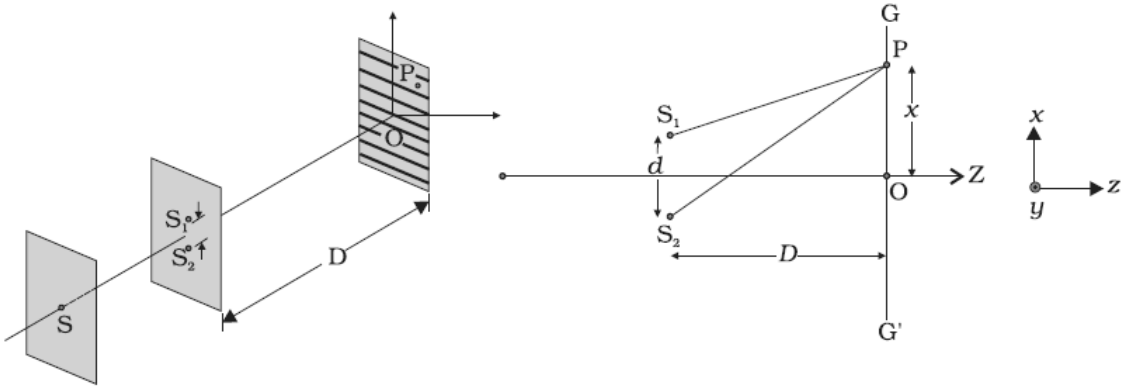


Fig. 3.12 Young's arrangement to produce interference pattern

Thus spherical waves emanating from S_1 and S_2 will produce interference fringes on the screen GG′, as shown in Fig. 3.12(b). The positions of maximum and minimum intensities can be calculated by using the analysis given in Section 10.4 where we had shown that for an arbitrary point P on the line GG′ [Fig. 3.12(b)] to correspond to a maximum, we must have

$$S_2P - S_1P = n\lambda; \qquad n = 0, 1, 2 \ldots \qquad (3.15)$$

Now,

$$(S_2P)^2 - (S_1P)^2 = \left[D^2 + \left(x + \frac{d}{2}\right)^2\right] - \left[D^2 + \left(x - \frac{d}{2}\right)^2\right] = 2xd$$

where S_1S_2 = d and OP = x .

Thus,

$$S_2P - S_1P = \frac{2xd}{S_2P + S_1P} \qquad (3.16)$$

If x, d<<D then negligible error will be introduced if $S_2P + S_1P$ (in the denominator) is replaced by 2D. For example, for d = 0.1 cm, D = 100 cm, OP = 1 cm (which correspond to typical values for an interference experiment using light waves), we have

$$S_2P - S_1P \approx [(100)^2 + (1.05)^2]^{\frac{1}{2}} + [(100)^2 + (0.95)^2]^{\frac{1}{2}}$$
$$\approx 200.01$$

Thus if we replace $S_2P + S_1P$ by 2 D, the error involved is about 0.005%. In this approximation, Eq. (3.16) becomes

$$S_2P - S_1P \approx \frac{xd}{D} \qquad (3.17)$$

Hence we will have constructive interference resulting in a bright region when $\frac{xd}{D} = n\lambda$ [Eq. (3.15)]. That is,

$$x = x_n = \frac{n\lambda D}{d}; n = 0, \pm 1, \pm 2, \ldots \qquad (3.18)$$

On the other hand, we will have destructive interference resulting in dark region when $\frac{xd}{D} = (n + \frac{1}{2})\lambda$ that is

$$x = x_n = (n + \frac{1}{2})\frac{\lambda D}{d}; n = 0, \pm 1, \pm 2, \ldots \qquad (3.19)$$

Thus dark and bright bands appear on the screen, as shown in Fig. 3.13. Such bands are called fringes. Equations (3.18) and (3.19) show that dark and bright fringes are equally spaced and the distance between two consecutive bright and dark fringes is given by

$$\beta = x_{n-1} - x_n \quad or \quad \beta = \frac{\lambda D}{d} \qquad (3.20)$$

which is the expression for the fringe width. Obviously, the central point O (in Fig. 3.12) will be bright because $S_1O = S_2O$ and it will correspond to n = 0 [Eq. (3.18)]. If we consider the line perpendicular to the plane of the paper and passing through O [i.e., along the y-axis] then all points on this line will be equidistant from S_1 and S_2 and we will have a bright central fringe which is a straight line as shown in Fig. 3.13. In order to determine the shape of the interference pattern on the screen we note that a particular fringe would correspond to the locus of points with a constant value of $S_2P - S_1P$. Whenever this constant is an integral multiple of λ, the fringe will be bright and whenever it is an odd integral multiple of λ/2 it will be a dark fringe. Now, the locus of the point P lying in the *x-y* plane such that $S_2P - S_1P$ (= Δ) is a constant, is a hyperbola. Thus the fringe pattern will strictly be a hyperbola; however, if the distance D is very large compared to the fringe width, the fringes will be very nearly straight lines as shown in Fig. 3.13

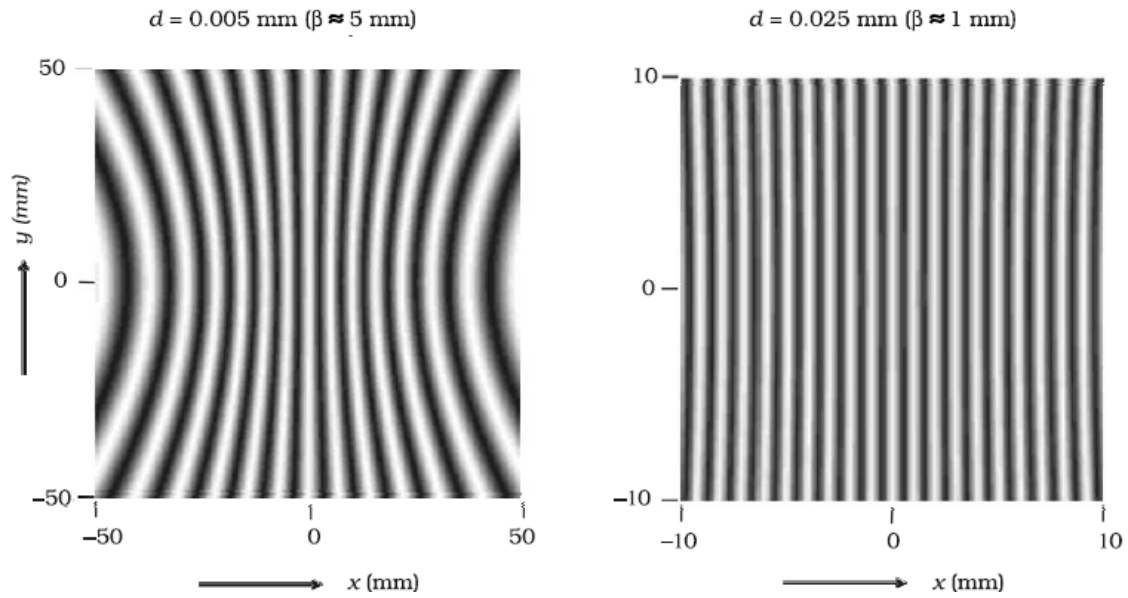
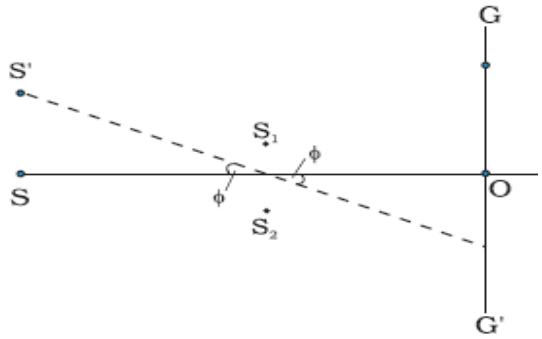


Fig. 3.13 Computer generated fringe pattern produced by two point source S_1 and S_2 on the screen GG' (3.12); (a) and (b) correspond to d = 0.005 mm and 0.025 mm. respectively (both figures correspond to D = 5 cm and $\lambda = 5 \times 10^{-5}$ cm.) (Adopted from OPTICS by A. Ghatak. Tata McGraw Hill publishing Co. Ltd., New Delhi, 2000.)

In the double-slit experiment shown in Fig. 3.12, we have taken the source hole S on the perpendicular bisector of the two slits, which is shown as the line SO. What happens if the source S is slightly away from the perpendicular bisector. Consider that the source is moved to some new point S′ and suppose that Q is the mid-point of S1 and S2. If the angle S′QS is φ, then



the central bright fringe occurs at an angle –φ, on the other side. Thus, if the source S is on the perpendicular bisector, then the central fringe occurs at O, also on the perpendicular bisector. If S is shifted by an angle φ to point S′, then the central fringe appears at a point O′ at an angle –φ, which means that it is shifted by the same angle on the other side of the bisector. This also means that the source S′, the mid-point Q and the point O′ of the central fringe are in a straight line.

We end this section by quoting from the Nobel lecture of Dennis Gabor*

---

*Dennis Gabor received the 1971 Nobel Prize in Physics for discovering the principles of holography

The wave nature of light was demonstrated convincingly for the first time in 1801 by Thomas Young by a wonderfully simple experiment. He let a ray of sunlight into a dark room, placed a dark screen in front of it, pierced with two small pinholes, and beyond this, at some distance, a white screen. He then saw two darkish lines at both sides of a bright line, which gave him sufficient encouragement to repeat the experiment, this time with spirit flame as light source, with a little salt in it to produce the bright yellow sodium light. This time he saw a number of dark lines, regularly spaced; the first clear proof that light added to light can produce darkness. This phenomenon is called interference. Thomas Young had expected it because he believed in the wave theory of light.

We should mention here that the fringes are straight lines although S_1 and S_2 are point sources. If we had slits instead of the point sources (Fig. 3.14), each pair of points would have produced straight line fringes resulting in straight line fringes with increased intensities.

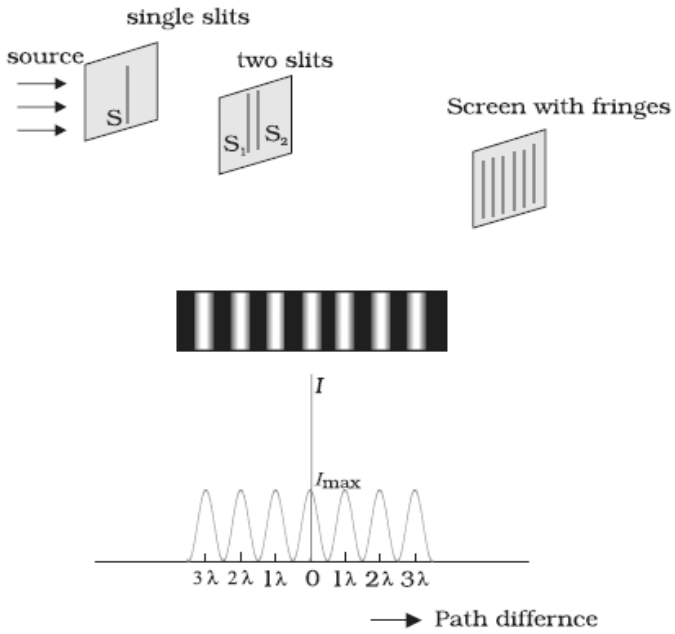


Fig. 3.14 Photograph and the graph of the intensity distribution in young's double-slit experiment.

3.6 DIFFRACTION

If we look clearly at the shadow cast by an opaque object, close to the region of geometrical shadow, there are altemate dark and bright regions just like in interference. This happens due to the phenomenon of diffraction. Diffraction is a general characteristic exhibited by all types of waves, be it sound waves, light waves, water waves or matter waves. Since the wavelength of light is much smaller than the dimensions of most obstacles; we do not encounter diffraction effects of light in everyday observations. However, the finite resolution of our eye or of optical instruments such as telescopes or microscopes is limited due to the phenomenon of diffraction. Indeed the colours that you see when a CD is viewed are due to diffraction effects. We will now discuss the phenomenon of diffraction.

3.6.1 The single slit

In the discussion of Young's experiment, we stated that a single narrow slit acts as a new source from which light spreads out. Even before Young, early experimenters – including Newton – had noticed that light spreads out from narrow holes and slits. It seems to turn around corners and enter regions where we would expect a shadow. These effects, known as diffraction, can only be properly understood using wave ideas. After all, you are hardly surprised to hear sound waves from someone talking around a corner!

When the double slit in Young's experiment is replaced by a single narrow slit (illuminated by monochromatic source), a broad pattern with a central bright region is seen. On both sides, there are alternate dark and bright regions, the intensity becoming weaker away from the centre (Fig. 3.16). To understand this, go to Fig. 3.15, which shows a parallel beam of light falling normally on a single slit LN of width a. The diffracted light goes on to meet a screen. The midpoint of the slit is M.

A straight line through M perpendicular to the slit plane meets the screen at C. We want the intensity at any point P on the screen. As before, straight lines joining P to the different points L,M,N, etc., can be treated as parallel, making an angle θ with the normal MC. The basic idea is to divide the slit into much smaller parts, and add their contributions at P with the proper phase differences. We are treating different parts of the wavefront at the slit as secondary sources. Because the incoming wavefront is parallel to the plane of the slit, these sources are in phase.

The path difference NP – LP between the two edges of the slit can be calculated exactly as for Young's experiment. From Fig. 3.15

$$\begin{aligned} NP - LP &= NQ \\ &= a \sin\theta \\ &= a\theta \quad \text{(for smaller angles)} \end{aligned} \qquad (3.21)$$

Similarly, if two points M_1 and M_2 in the slit plane are separated by y, the path difference $M_2P - M_1P \approx y\theta$. We now have to sum up equal, coherent contributions from a large number of sources, each with a different phase. This calculation was made by Fresnel using integral calculus, so we omit it here. The main features of the diffraction pattern can be understood by simple arguments.

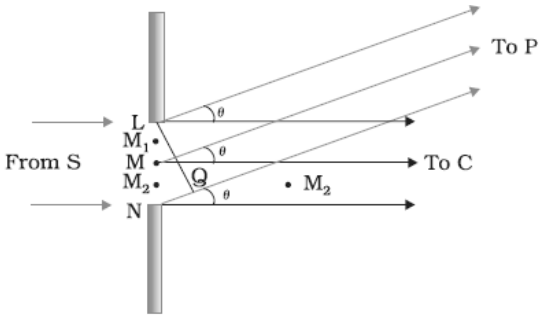


Fig. 3.15 The geometry of path differences for diffraction by single slit.

At the central point C on the screen, the angle θ is zero. All path differences are zero and hence all the parts of the slit contribute in phase. This gives maximum intensity at C. Experimental observation shown in Fig. 3.15 indicates that the intensity has a central maximum at $\theta = 0$ and other secondary maxima at $\theta \approx (n+1/2)\,\lambda/a$, and has minima (zero intensity) at $\theta \approx$

$n\lambda/a$, n = ±1, ±2, ±3, ... It is easy to see why it has minima at these values of angle. Consider first the angle θ where the path difference aθ is λ. Then

$$\theta = \frac{\lambda}{a} \qquad (3.22)$$

Now, divide the slit into two equal halves LM and MN each of size $a/2$. For every point M_1 in LM, there is a point M_2 in MN such that $M_1M_2 = a/2$. The path difference between M1 and M_2 at $P = M_2P - M_1P = \theta\, a\, /2 = \lambda/2$ for the angle chosen. This means that the contributions from M_1 and M_2 are 180° out of phase and cancel in the direction $\theta = \lambda/a$. Contributions from the two halves of the slit LM and MN, therefore, cancel each other. Equation (3.22) gives the angle at which the intensity falls to zero. One can similarly show that the intensity is zero for $\theta = n\lambda/a$, with n being any integer (except zero!). Notice that the angular size of the central maximum increases when the slit width a decreases.

It is also easy to see why there are maxima at $\theta = (n + 1/2)\, \lambda/a$ and why they go on becoming weaker and weaker with increasing n. Consider an angle $\theta = 3\lambda/2a$ which is midway between two of the dark fringes. Divide the slit into three equal parts. If we take the first two thirds of the slit, the path difference between the two ends would be

$$\frac{2}{3}a \times \theta = \frac{2a}{3} \times \frac{3\lambda}{2a} = \lambda \qquad (3.23)$$

The first two-thirds of the slit can therefore be divided into two halves which have a λ/2 path difference. The contributions of these two halves cancel in the same manner as described earlier. Only the remaining one-third of the slit contributes to the intensity at a point between the two minima. Clearly, this will be much weaker than the central maximum (where the entire slit contributes in phase). One can similarly show that there are maxima at $(n + 1/2)\, \theta/a$ with n = 2, 3, etc. These become weaker with increasing n, since only one-fifth, one-seventh, etc., of the slit contributes in these cases. The photograph and intensity pattern corresponding to it is shown in Fig. 3.16.

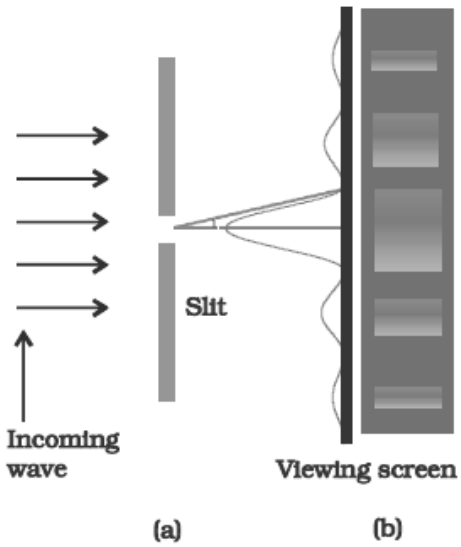


Fig. 3.16 Intensity distribution and photograph of fringes due to diffraction at single slit.

There has been prolonged discussion about difference between intereference and diffraction among scientists since the discovery of these phenomena. In this context, it is interesting to note what Richard Feynman* has said in his famous Feynman Lectures on Physics:

No one has ever been able to define the difference between interference and diffraction satisfactorily. It is just a question of usage, and there is no specific, important physical difference between them. The best we can do is, roughly speaking, is to say that when there are only a few sources, say two interfering sources, then the result is usually called interference, but if there is a large number of them, it seems that the word diffraction is more often used.

In the double-slit experiment, we must note that the pattern on the screen is actually a superposition of single-slit diffraction from each slit or hole, and the double-slit interference pattern. This is shown in Fig. 3.17. It shows a broader diffraction peak in which there appear several fringes of smaller width due to double-slit interference. The number of interference fringes occuring in the broad diffraction peak depends on the ratio d/a, that is the ratio of the distance between the two slits to the width of a slit. In the limit of a becoming very small, the diffraction pattern will become very flat and we will observe the two-slit interference pattern [see Fig. 3.13(b)]

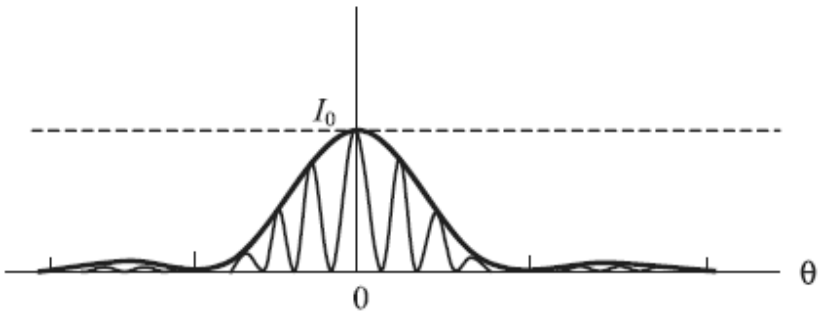


Fig. 3.17 The actual double-slit interference pattern. The envelope shows the single slit diffraction.

In the double-slit interference experiment of Fig. 3.12, what happens if we close one slit? You will see that it now amounts to a single slit. But you will have to take care of some shift in the pattern. We now have a source at S, and only one hole (or slit) S_1 or S_2. This will produce a single- slit diffraction pattern on the screen. The centre of the central bright fringe will appear at a point which lies on the straight line SS_1 or SS_2, as the case may be.

We now compare and contrast the interference pattern with that seen for a coherently illuminated single slit (usually called the single slit diffraction pattern).

(i) The interference pattern has a number of equally spaced bright and dark bands. The diffraction pattern has a central bright maximum which is twice as wide as the other maxima. The intensity falls as we go to successive maxima away from the centre, on either side.

(ii) We calculate the interference pattern by superposing two waves originating from the two narrow slits. The diffraction pattern is a superposition of a continuous family of waves originating from each point on a single slit.

---

* Richard Feynman was one of the recipients of the 1965 Nobel Prize in Physica for his fundsmental work in quantum electrodynamics

(iii) For a single slit of width a, the first null of the interference pattern occurs at an angle of λ/a. At the same angle of λ/a, we get a maximum (not a null) for two narrow slits separated by a distance a.

One must understand that both d and a have to be quite small, to be able to observe good interference and diffraction patterns. For example, the separation d between the two slits must be of the order of a millimeter or so. The width a of each slit must be even smaller, of the order of 0.1 or 0.2 mm.

In our discussion of Young's experiment and the single-slit diffraction, we have assumed that the screen on which the fringes are formed is at a large distance. The two or more paths from the slits to the screen were treated as parallel. This situation also occurs when we place a converging lens after the slits and place the screen at the focus. Parallel paths from the slit are combined at a single point on the screen. Note that the lens does not introduce any extra path differences in a parallel beam. This arrangement is often used since it gives more intensity than placing the screen far away. If f is the focal length of the lens, then we can easily work out the size of the central bright maximum. In terms of angles, the separation of the central maximum from the first null of the diffraction pattern is λ/a. Hence, the size on the screen will be f λ/a.

3.6.2 Seeing the single slit diffraction pattern

It is surprisingly easy to see the single-slit diffraction pattern for oneself. The equipment needed can be found in most homes — two razor blades and one clear glass electric bulb preferably with a straight filament. One has to hold the two blades so that the edges are parallel and have a narrow slit in between. This is easily done with the thumb and forefingers (Fig. 3.18).

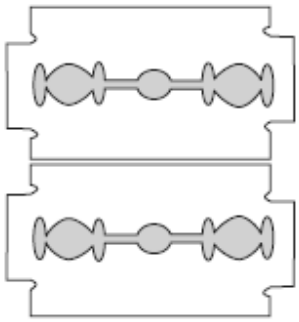


Fig. 3.18 Holding two blades to form a single slit. A blub filament viewed through this shows clear diffraction bands.

Keep the slit parallel to the filament, right in front of the eye. Use spectacles if you normally do. With slight adjustment of the width of the slit and the parallelism of the edges, the pattern should be seen with its bright and dark bands. Since the position of all the bands (except the central one) depends on wavelength, they will show some colours. Using a filter for red or blue will make the fringes clearer. With both filters available, the wider fringes for red compared to blue can be seen.

In this experiment, the filament plays the role of the first slit S in Fig. 10.16. The lens of the eye focuses the pattern on the screen (the retina of the eye). With some effort, one can cut a double slit in an aluminium foil with a blade. The bulb filament can be viewed as before to repeat Young's experiment. In daytime, there is another suitable bright source subtending a small angle at the eye. This is the reflection of the Sun in any shiny convex surface (e.g., a cycle bell).

Do not try direct sunlight – it can damage the eye and will not give fringes anyway as the Sun subtends an angle of (1/2)º.

In interference and diffraction, light energy is redistributed. If it reduces in one region, producing a dark fringe, it increases in another region, producing a bright fringe. There is no gain or loss of energy, which is consistent with the principle of conservation of energy.

3.6.3 Resolving power of optical instruments

We had discussed about telescopes. The angular resolution of the telescope is determined by the objective of the telescope. The stars which are not resolved in the image produced by the objective cannot be resolved by any further magnification produced by the eyepiece. The primary purpose of the eyepiece is to provide magnification of the image produced by the objective.

Consider a parallel beam of light falling on a convex lens. If the lens is well corrected for aberrations, then geometrical optics tells us that the beam will get focused to a point. However, because of diffraction, the beam instead of getting focused to a point gets focused to a spot of finite area. In this case the effects due to diffraction can be taken into account by considering a plane wave incident on a circular aperture followed by a convex lens (Fig. 3.19). The analysis of the corresponding diffraction pattern is quite involved; however, in principle, it is similar to the analysis carried out to obtain the single-slit diffraction pattern. Taking into account the effects due to diffraction, the pattern on the focal plane would consist of a central bright region surrounded by concentric dark and bright rings (Fig. 3.19). A detailed analysis shows that the radius of the central bright region is approximately given by

$$r_0 \approx \frac{1.22\lambda f}{2a} = \frac{0.61\lambda f}{a} \qquad (3.24)$$

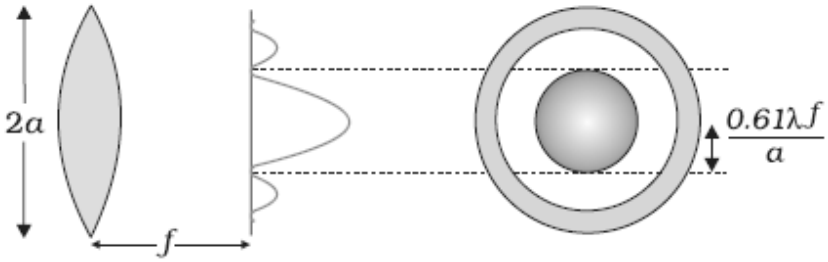


Fig. 3.19 A parallel beam of light is incident on a convex lens.

Because of diffraction effects. The beam gets focused to a spot of radius $\approx 0.61\frac{\lambda f}{a}$.

where f is the focal length of the lens and 2a is the diameter of the circular aperture or the diameter of the lens, whichever is smaller. Typically if

$$\lambda \approx 0.5\ \mu m,\ f \approx 20\ cm \text{ and } a \approx 5\ cm$$

we have,

$$r_0 \approx 1.2\ \mu m$$

Although the size of the spot is very small, it plays an important role in determining the limit of resolution of optical instruments like a telescope or a microscope. For the two stars to be just resolved

$$f\Delta\theta \approx r_0 \approx \frac{0.61\lambda f}{a}$$

Implying

$$\Delta\theta \approx \frac{0.61\lambda}{a} \qquad (3.25)$$

Thus, Δθ will be small if the diameter of the objective is large. This implies that the telescope will have better resolving power if a is large. It is for this reason that for better resolution, a telescope must have a large diameter objective.

We can apply a similar argument to the objective lens of a microscope. In this case, the object is placed slightly beyond f, so that a real image is formed at a distance v [Fig. 3.20]. The magnification – ratio of image size to object size – is given by m ≈v/f. It can be seen from Fig. 3.20 that

$$\frac{D}{f} \approx 2tan\beta \qquad (3.26)$$

where, 2β is the angle subtended by the diameter of the objective lens at the focus of the microscope.

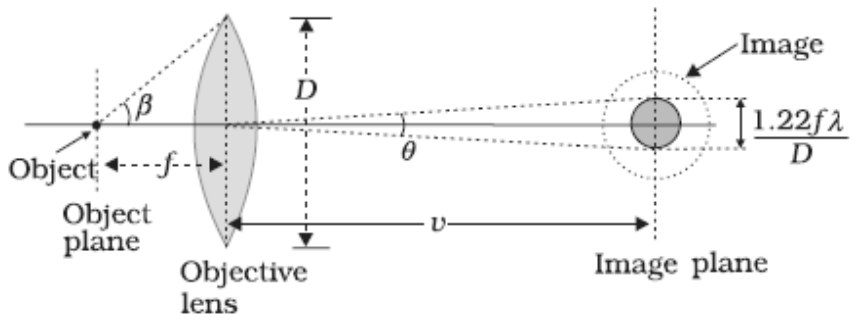


Fig. 3.20 Real image formed by the objective lens of the microscope.

DETERMINE THE RESOLVING POWER OF YOUR EYE

You can estimate the resolving power of your eye with a simple experiment. Make black stripes of equal width separated by white stripes; see figure here. All the black stripes should be of equal width, while the width of the intermediate white stripes should increase as you go from the left to the right. For example, let all black stripes have a width of 5 mm. Let the width of the first two white stripes be 0.5 mm each, the next two white stripes be 1 mm each, the next two 1.5 mm each, etc. Paste this pattern on a wall in a room or laboratory, at the height of your eye.

Now watch the pattern, preferably with one eye. By moving away or closer to the wall, find the position where you can just see some two black stripes as separate stripes. All the black stripes to the left of this stripe would merge into one another and would not be distinguishable. On the other hand, the black stripes to the right of this would be more and more clearly visible. Note the width d of the white stripe which separates the two regions, and measure the distance D of the wall from your eye. Then d/D is the resolution of your eye.

You have watched specks of dust floating in air in a sunbeam entering through your window. Find the distance (of a speck) which you can clearly see and distinguish from a

neighbouring speck. Knowing the resolution of your eye and the distance of the speck, estimate the size of the speck of dust.

When the separation between two points in a microscopic specimen is comparable to the wavelength λ of the light, the diffraction effects become important. The image of a point object will again be a diffraction pattern whose size in the image plane will be

$$\vartheta\theta = \vartheta\left(\frac{1.22\lambda}{D}\right) \qquad (3.27)$$

Two objects whose images are closer than this distance will not be resolved, they will be seen as one. The corresponding minimum separation, d_{min}, in the object plane is given by

$$d_{min} = \frac{\left[\vartheta\left(\frac{1.22\lambda}{D}\right)\right]}{m}$$

$$= \frac{1.22\lambda}{D}.\frac{v}{m}$$

Or,

since m= $\frac{v}{f}$

$$= \frac{1.22f\lambda}{D} \qquad (3.28)$$

Now, combining Eqs. (3.26) and (3.28), we get

$$d_{mim} = \frac{1.22\ \lambda}{2\ tan\beta}$$

$$= \frac{1.22\ \lambda}{2\ sin\beta} \qquad (3.29)$$

If the medium between the object and the objective lens is not air but a medium of refractive index n, Eq. (3.29) gets modified to

$$d_{mim} = \frac{1.22\ \lambda}{2n\ sin\beta} \qquad (3.30)$$

The product n sinβ is called the numerical aperture and is sometimes marked on the objective.

The resolving power of the microscope is given by the reciprocal of the minimum separation of two points seen as distinct. It can be seen from Eq. (3.30) that the resolving power can be increased by choosing a medium of higher refractive index. Usually an oil having a refractive index close to that of the objective glass is used. Such an arrangement is called an 'oil immersion objective'. Notice that it is not possible to make sinβ larger than unity. Thus, we see that the resolving power of a microscope is basically determined by the wavelength of the light used.

There is a likelihood of confusion between resolution and magnification, and similarly between the role of a telescope and a microscope to deal with these parameters. A telescope produces images of far objects nearer to our eye. Therefore objects which are not resolved at far distance can be resolved by looking at them through a telescope. A microscope, on the other hand, magnifies objects (which are near to us) and produces their larger image. We may be looking at two stars or two satellites of a far-away planet, or we may be looking at different regions of a living cell. In this context, it is good to remember that a telescope resolves whereas a microscope magnifies.

10.6.4 The validity of ray optics

An aperture (i.e., slit or hole) of size a illuminated by a parallel beam sends diffracted light into an angle of approximately $\approx \lambda/a$. This is the angular size of the bright central maximum. In travelling a distance z, the diffracted beam therefore acquires a width $z\lambda/a$ due to diffraction. It is interesting to ask at what value of z the spreading due to diffraction becomes comparable to the size a of the aperture. We thus approximately equate $z\lambda/a$ with a. This gives the distance beyond which divergence of the beam of width a becomes significant. Therefore,

$$z \cong \frac{a^2}{\lambda} \qquad (3.31)$$

We define a quantity z_F called the Fresnel distance by the following equation

$$z_F \cong \frac{a^2}{\lambda}$$

Equation (3.31) shows that for distances much smaller than Z_F, the spreading due to diffraction is smaller compared to the size of the beam. It becomes comparable when the distance is approximately Z_F. For distances much greater than Z_F, the spreading due to diffraction dominates over that due to ray optics (i.e., the size a of the aperture). Equation (3.31) also shows that ray optics is valid in the limit of wavelength tending to zero.

10.7 POLARISATION

Consider holding a long string that is held horizontally, the other end of which is assumed to be fixed. If we move the end of the string up and down in a periodic manner, we will generate a wave propagating in the +x direction (Fig. 3.22). Such a wave could be described by the following equation

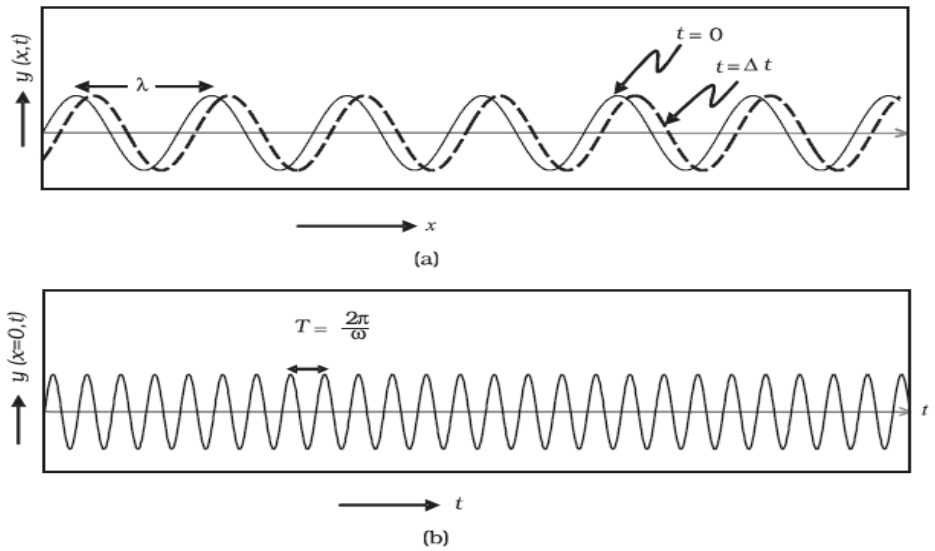


Fig. 3.21(a) The curve represent the displacement of a string at t=0 and t=Δt, respectively when a sinusoidal wave is propagating in +*x*-direction.

(b) The curve represents the time variation of the displacement at *x*=0 when a sinusoidal wave is propagating in +x- direction. At *x*=Δ*x* the time variation of the displacement will be slightly displaced to the right.

$$y(x,t) = a \sin(kx - \omega t) \qquad (3.32)$$

where a and ω ($= 2\pi\nu$) represent the amplitude and the angular frequency of the wave, respectively; further,

$$\lambda = \frac{2\pi}{k} \qquad (3.33)$$

represents the wavelength associated with the wave. We had discussed propagation of such waves in Chapter 15 of Class XI textbook. Since the displacement (which is along the y direction) is at right angles to the direction of propagation of the wave, we have what is known as a transverse wave. Also, since the displacement is in the y direction, it is often referred to as a y-polarised wave. Since each point on the string moves on a straight line, the wave is also referred to as a linearly polarized wave. Further, the string always remains confined to the x-y plane and therefore it is also referred to as a plane polarised wave. In a similar manner we can consider the vibration of the string in the *x*-*z* plane generating a z-polarised wave whose displacement will be given by

$$z(x,t) = a \sin(kx - \omega t) \qquad (3.34)$$

It should be mentioned that the linearly polarised waves [described by Eqs. (3.33) and (3.34)] are all transverse waves; i.e., the displacement of each point of the string is always at right angles to the direction of propagation of the wave. Finally, if the plane of vibration of the string is changed randomly in very short intervals of time, then we have what is known as an unpolarised wave. Thus, for an unpolarised wave the displacement will be randomly changing with time though it will always be perpendicular to the direction of propagation.

Light waves are transverse in nature; i.e., the electric field associated with a propagating light wave is always at right angles to the direction of propagation of the wave. This can be easily demonstrated using a simple polaroid. You must have seen thin plastic like sheets, which are called polaroids. A polaroid consists of long chain molecules aligned in a particular direction. The electric vectors (associated with the propagating light wave) along the direction of the aligned molecules get absorbed. Thus, if an unpolarised light wave is incident on such a polaroid then the light wave will get linearly polarised with the electric vector oscillating along a direction perpendicular to the aligned molecules; this direction is known as the pass-axis of the polaroid.

Thus, if the light from an ordinary source (like a sodium lamp) passes through a polaroid sheet P_1, it is observed that its intensity is reduced by half. Rotating P1 has no effect on the transmitted beam and transmitted intensity remains constant. Now, let an identical piece of polaroid P_2 be placed before P_1. As expected, the light from the lamp is reduced in intensity on passing through P_2 alone. But now rotating P_1 has a dramatic effect on the light coming from P_2. In one position, the intensity transmitted by P_2 followed by P_1 is nearly zero. When turned by 90º from this position, P_1 transmits nearly the full intensity emerging from P_2 (Fig. 3.22).

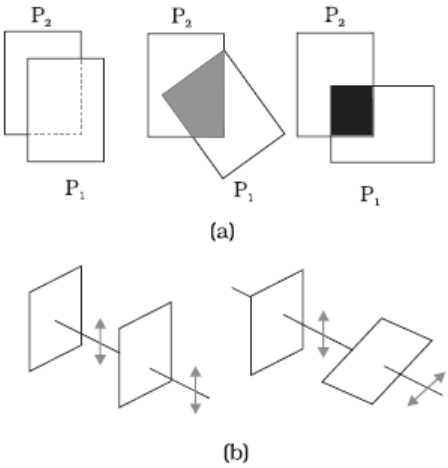


Fig. 3.22 (a) Passage of light through two polaroids P_2 and P_1. The transmitted fraction falls from 1 to 0 as the angle between them varies from 0º to 90º. Notice that the light seen through a single Polaroid P_1 does not vary with angle.

(b) Behaviour of the electric vector when light passes through two polaroids. The transmitted polarisation is the component parallel to the polaroid axis. The double arrows show the oscillations of the electric vector.

The above experiment can be easily understood by assuming that light passing through the polaroid P_2 gets polarised along the pass-axis of P_2. If the pass-axis of P_2 makes an angle θ with the pass-axis of P1, then when the polarised beam passes through the polaroid P_2, the component E cos θ (along the pass-axis of P_2) will pass through P_2. Thus, as we rotate the polaroid P_1 (or P_2), the intensity will vary as:

$$I = I_0 \cos 2\theta \qquad (3.35)$$

where I_0 is the intensity of the polarized light after passing through P1. This is known as Malus' law. The above discussion shows that the intensity coming out of a single polaroid is half of the incident intensity. By putting a second polaroid, the intensity can be further controlled from 50% to zero of the incident intensity by adjusting the angle between the pass-axes of two polaroids.

Polaroids can be used to control the intensity, in sunglasses, windowpanes, etc. Polaroids are also used in photographic cameras and 3D movie cameras.

10.7.1 Polarisation by scattering

The light from a clear blue portion of the sky shows a rise and fall of intensity when viewed through a polaroid which is rotated. This is nothing but sunlight, which has changed its direction (having been scattered) on encountering the molecules of the earth's atmosphere. As Fig. 3.24(a) shows, the incident sunlight is unpolarised. The dots stand for polarization perpendicular to the plane of the figure. The double arrows show polarisation in the plane of the figure. (There is no phase relation between these two in unpolarised light). Under the influence of the electric field of the incident wave the electrons in the molecules acquire components of motion in both these directions. We have drawn an observer looking at 90° to the direction of the sun. Clearly, charges accelerating parallel to the double arrows do not radiate energy towards this observer since their acceleration has no transverse component. The radiation scattered by the

molecule is therefore represented by dots. It is polarized perpendicular to the plane of the figure. This explains the polarisation of scattered light from the sky.

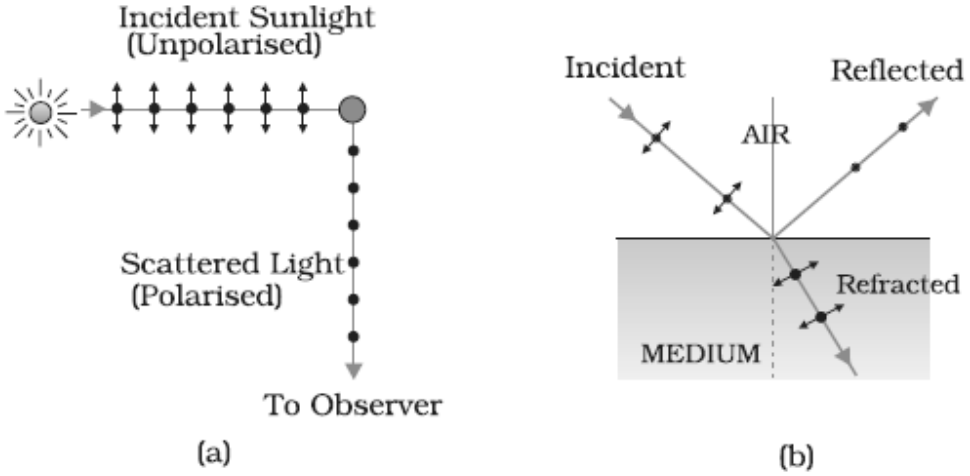


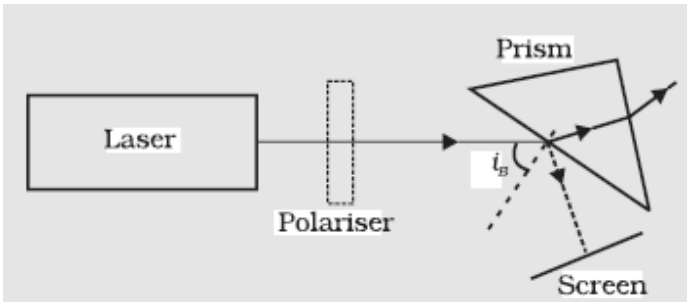
Fig. 3.23 (a) Polarisation of the blue scattered light from the sky. The incident sunlight is unpolarised (dots and arrows). A typical molecule is shown. It scatters light by 90º polarised normal to the plane of the paper (dots only).

(b) Polarisation of lightreflected from a transparent medium at the Brewster angle (reflected ray perpendicular to refracted ray).

The scattering of light by molecules was intensively investigated by C.V. Raman and his collaborators in Kolkata in the 1920s. Raman was awarded the Nobel Prize for Physics in 1930 for this work.

When light is incident on an interface of two media, it is observed that some part of it gets reflected and some part gets transmitted. Consider a related question:

Is it possible that under some conditions a monochromatic beam of light incident on a surface (which is normally reflective) gets completely transmitted with no reflection? To your surprise, the answer is *yes*.



Let us try a simple experiment and check what happens. Arrange a laser, a good polariser, a prism and screen as shown in the figure here. Let the light emitted by the laser source pass through the polariser and be incident on the surface of the prism at the Brewster's angle of incidence *i*B. Now rotate the polariser carefully and you will observe that for a specific alignment of the polariser, the light incident on the prism is completely transmitted and no light is reflected from the surface of the prism. The reflected spot will completely vanish.

10.7.2 Polarisation by reflection

Figure 3.23(b) shows light reflected from a transparent medium, say, water. As before, the dots and arrows indicate that both polarisations are present in the incident and refracted waves. We have drawn a situation in which the reflected wave travels at right angles to the refracted wave. The oscillating electrons in the water produce the reflected wave. These move in the two directions transverse to the radiation from wave in the medium, i.e., the refracted wave. The arrows are parallel to the direction of the reflected wave. Motion in this direction does not contribute to the reflected wave. As the figure shows, the reflected light is therefore linearly polarised perpendicular to the plane of the figure (represented by dots). This can be checked by looking at the reflected light through an analyser. The transmitted intensity will be zero when the axis of the analyser is in the plane of the figure, i.e., the plane of incidence. When unpolarised light is incident on the boundary between two transparent media, the reflected light is polarised with its electric vector perpendicular to the plane of incidence when the refracted and reflected rays make a right angle with each other. Thus we have seen that when reflected wave is perpendicular to the refracted wave, the reflected wave is a totally polarised wave. The angle of incidence in this case is called Brewster's angle and is denoted by iB. We can see that i_B is related to the refractive index of the denser medium. Since we have $i_B + r = \pi/2$, we get from Snell's law

$$\mu = \frac{sin\, i_B}{\sin r} = \frac{\sin i_B}{\sin\left(i_B - \frac{\pi}{2}\right)}$$

$$= \frac{\sin i_B}{\cos i_B} = \tan i_B \qquad (3.36)$$

This is known as Brewster's law.

For simplicity, we have discussed scattering of light by 90°, and reflection at the Brewster angle. In this special situation, one of the two perpendicular components of the electric field is zero. At other angles, both components are present but one is stronger than the other. There is no stable phase relationship between the two perpendicular components since these are derived from two perpendicular components of an unpolarised beam. When such light is viewed through a rotating analyser, one sees a maximum and a minimum of intensity but not complete darkness. This kind of light is called partially polarised.

Let us try to understand the situation. When an unpolarised beam of light is incident at the Brewster's angle on an interface of two media, only part of light with electric field vector perpendicular to the plane of incidence will be reflected. Now by using a good polariser, if we completely remove all the light with its electric vector perpendicular to the plane of incidence and let this light be incident on the surface of the prism at Brewster's angle, you will then observe no reflection and there will be total transmission of light.

SUMMARY

1. Huygens' principle tells us that each point on a wavefront is a source of secondary waves, which add up to give the wavefront at a later time.
2. Huygens' construction tells us that the new wavefront is the forward envelope of the secondary waves. When the speed of light is independent of direction, the secondary waves are spherical. The rays are then perpendicular to both the wavefronts and the time of travelis the same measured along any ray. This principle leads to the well known laws of reflection and refraction.

3. The principle of superposition of waves applies whenever two or more sources of light illuminate the same point. When we consider the intensity of light due to these sources at the given point, there is an interference term in addition to the sum of the individual intensities. But this term is important only if it has a non-zero average, which occurs only if the sources have the same frequency and a stable phase difference.
4. Young's double slit of separation d gives equally spaced fringes of angular separation λ/d. The source, mid-point of the slits, and central bright fringe lie in a straight line. An extended source will destroy the fringes if it subtends angle more than λ/d at the slits.
5. A single slit of width a gives a diffraction pattern with a central maximum. The intensity falls to zero at angles of $\pm\frac{\lambda}{a}, \pm\frac{2\lambda}{a}$, etc., with successively weaker secondary maxima in between. Diffraction limits the angular resolution of a telescope to λ/D where D is the diameter. Two stars closer than this give strongly overlapping images. Similarly, a microscope objective subtending angle 2β at the focus, in a medium of refractive index n, will just separate two objects spaced at a distance λ/(2n sin β), which is the resolution limit of a microscope. Diffraction determines the limitations of the concept of light rays. A beam of width a travels a distance $a^{2/}\lambda$, called the Fresnel distance, before it starts to spread out due to diffraction.
6. Natural light, e.g., from the sun is unpolarised. This means the electric vector takes all possible directions in the transverse plane, rapidly and randomly, during a measurement. A polaroid transmits only one component (parallel to a special axis). The resulting light is called linearly polarised or plane polarised. When this kind of light is viewed through a second polaroid whose axis turns through 2π, two maxima and minima of intensity are seen. Polarised light can also be produced by reflection at a special angle (called the Brewster angle) and by scattering through π/2 in the earth's atmosphere.

VERY SHORT ANSWER QUESTIONS (2 MARKS)

1. What is Fresnel distance?
2. Give the justification for validity of ray optics.
3. What is polarisation of light?
4. What is Mauls' law?
5. Explain Brewster's law.
6. When does a monochromatic beam of light incident on a reflective surface get completely transmitted?

SHORT ANSWER QUESTIONS (4 MARKS)

1. Explain Doppler effect in light. Distinguish between reds shift and blue shift.
2. Derive the expression for the intensity at a point where interference of light occurs. Arrive at the conditions for maximum and zero intensity.
3. Does the principal of conservation of energy hold for interference and diffraction phenomena ? Explain briefly.
4. How do you determine the resolving power of your eye?

LONG ANSWER QUESTIONS (8 MARKS)

1. Distinguish between Coherent and Incoherent addition of waves. Develop the theory of constructive and destructive interferences.
2. Describe Young's experiment for observing interference and hence arrive at the expression for 'fringe width'.
3. What is diffraction? Discuss diffraction pattern obtainable from a single slit.
4. What is resolving power of Optical Instruments? Derive the condition under which images are resolved.

CHAPTER 4

ELECTRIC CHARG ES AND FIELDS

4.1 INTRODUCTION

All of us have the experience of seeing a spark or hearing a crackle when we take off our synthetic clothes or sweater, particularly in dry weather. This is almost inevitable with ladies garments like a polyester saree. Have you ever tried to find any explanation for this phenomenon? Another common example of electric discharge is the lightning that we see in the sky during thunderstorms. We also experience a sensation of an electric shock either while opening the door of a car or holding the iron bar of a bus after sliding from our seat. The reason for these experiences is discharge of electric charges through our body, which were accumulated due to rubbing of insulating surfaces. You might have also heard that this is due to generation of static electricity. This is precisely the topic we are going to discuss in this and the next chapter. Static means anything that does not move or change with time. *Electrostatics deals with the study of forces, fields and potentials arising from static charges.*

4.2 ELECTRIC CHARGE

Historically the credit of discovery of the fact that amber rubbed with wool or silk cloth attracts light objects goes to Thales of Miletus, Greece, around 600 BC. The name electricity is coined from the Greek word elektron meaning amber. Many such pairs of materials were known which on rubbing could attract light objects like straw, pith balls and bits of papers. You can perform the following activity at home to experience such an effect. Cut out long thin strips of white paper and lightly iron them. Take them near a TV screen or computer monitor. You will see that the strips get attracted to the screen. In fact they remain stuck to the screen for a while.

It was observed that if two glass rods rubbed with wool or silk cloth are brought close to each other, they repel each other [Fig. 4.1(a)]. The two strands of wool or two pieces of silk cloth, with which the rods were rubbed, also repel each other. However, the glass rod and wool attracted each other. Similarly, two plastic rods rubbed with cat’s fur repelled each other [Fig.

4.1(b)] but attracted the fur. On the other hand, the plastic rod attracts the glass rod [Fig. 4.1 (c)] and repel the silk or wool with which the glass rod is rubbed. The glass rod repels the fur.

If a plastic rod rubbed with fur is made to touch two small pith balls(now-a-days we can use polystyrene balls) suspended by silk or nylon thread, then the balls repel each other [Fig. 4.1(d)] and are also repelled by the rod. A similar effect is found if the pith balls are touched with a glass rod rubbed with silk [Fig. 4.1(e)]. A dramatic observation is that a pith ball touched with glass rod attracts another pith ball touched with plastic rod [Fig. 4.1 (f)].

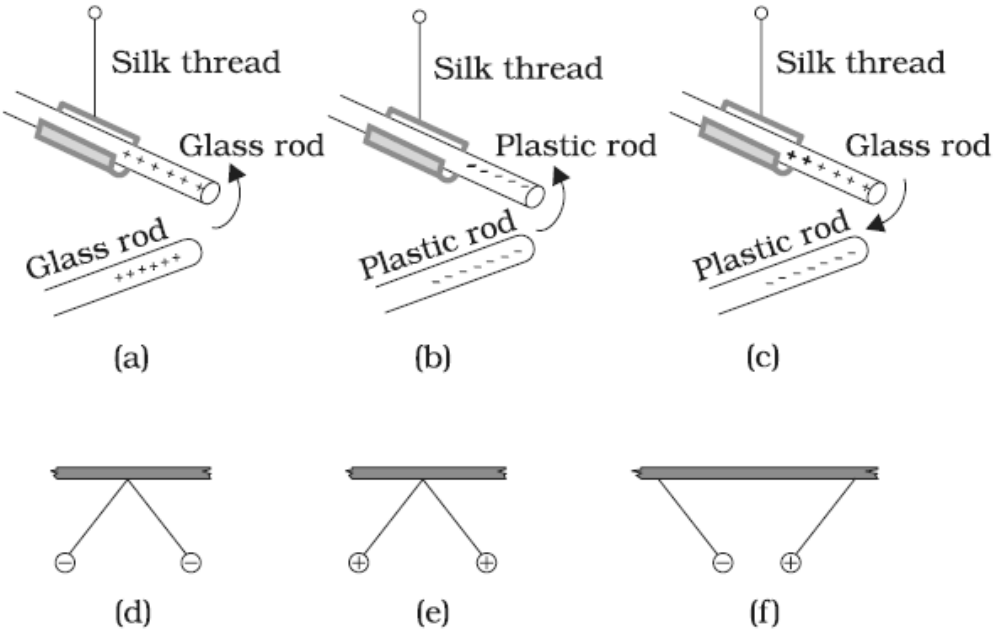


Fig. 4.1 Rods and pith balls: like charges repel and unlike charges attract each other.

These seemingly simple facts were established from years of efforts and careful experiments and their analyses. It was concluded, after many careful studies by different scientists, that there were only two kinds of an entity which is called the electric charge. We say that the bodies like glass or plastic rods, silk, fur and pith balls are electrified. They acquire an *electric charge* on rubbing. The experiments on pith balls suggested that there are two kinds of electrification and we find that (i) *like charges repel* and (ii) *unlike charges attract* each other. The experiments also demonstrated that the charges are transferred from the rods to the pith balls on contact. It is said that the pith balls are electrified or are charged by contact. The property which differentiates the two kinds of charges is called the polarity of charge.

When a glass rod is rubbed with silk, the rod acquires one kind of charge and the silk acquires the second kind of charge. This is true for any

pair of objects that are rubbed to be electrified. Now if the electrified glass rod is brought in contact with silk, with which it was rubbed, they no longer attract each other. They also do not attract or repel other light objects as they did on being electrified.

Thus, the charges acquired after rubbing are lost when the charged bodies are brought in contact. What can you conclude from these observations? It just tells us that unlike charges acquired by the objects neutralise or nullify each other's effect. Therefore the charges were named as *positive* and *negative* by the American scientist Benjamin Franklin. We know that when we add a positive number to a negative number of the same magnitude, the sum is zero. This might have been the philosophy in naming the charges as positive and negative. By convention, the charge on glass rod or cat's fur is called positive and that on plastic rod or silk is termed negative. If an object possesses an electric charge, it is said to be electrified or charged. When it has no charge it is said to be neutral.

UNIFICATION OF ELECTRICITY AND MAGNETISM

In olden days, electricity and magnetism were treated as separate subjects. Electricity dealt with charges on glass rods, cat's fur, batteries, lightning, etc., while magnetism described interactions of magnets, iron filings, compass needles, etc. In 1820 Danish scientist Oersted found that a compass needle is deflected by passing an electric current through a wire placed near the needle. Ampere and Faraday supported this observation by saying that electric charges in motion produce magnetic fields and moving magnets generate electricity. The unification was achieved when the Scottish physicist Maxwell and the Dutch physicist Lorentz put forward a theory where they showed the interdependence of these two subjects. This field is called electromagnetism. Most of the phenomena occurring around us can be described under electromagnetism. Virtually every force that we can think of like friction, chemical force between atoms holding the matter together, and even the forces describing processes occurring in cells of living organisms, have its origin in electromagnetic force. Electromagnetic force is one of the fundamental forces of nature.

Maxwell put forth four equations that play the same role in classical electromagnetism as Newton's equations of motion and gravitation law play in mechanics. He also argued that light is electromagnetic in nature and its speed can be found by making purely electric and magnetic measurements. He claimed that the science of optics is intimately related to that of electricity and magnetism.

The science of electricity and magnetism is the foundation for the modern technological civilisation. Electric power, telecommunication, radio and television, and a wide variety of the practical appliances used in daily life are based on the principles of this science. Although charged particles in motion exert both electric and magnetic forces, in the frame of reference where all the charges are at rest, the forces are purely electrical. You know that gravitational force is a long-range force. Its effect is felt even when the distance between the interacting particles is very large because the force decreases inversely as the square of the distance between the interacting bodies. We will learn in this chapter that electric force is also as pervasive and is in fact stronger than the gravitational force by several orders of

magnitude (refer to Chapter 1 of Class XI Physics Textbook).

A simple apparatus to detect charge on a body is the gold-leaf electroscope [Fig. 4.2(a)]. It consists of a vertical metal rod housed in a box, with two thin gold leaves attached to its bottom end. When a charged object touches the metal knob at the top of the rod, charge flows on to the leaves and they diverge. The degree of divergance is an indicator of the amount of charge.

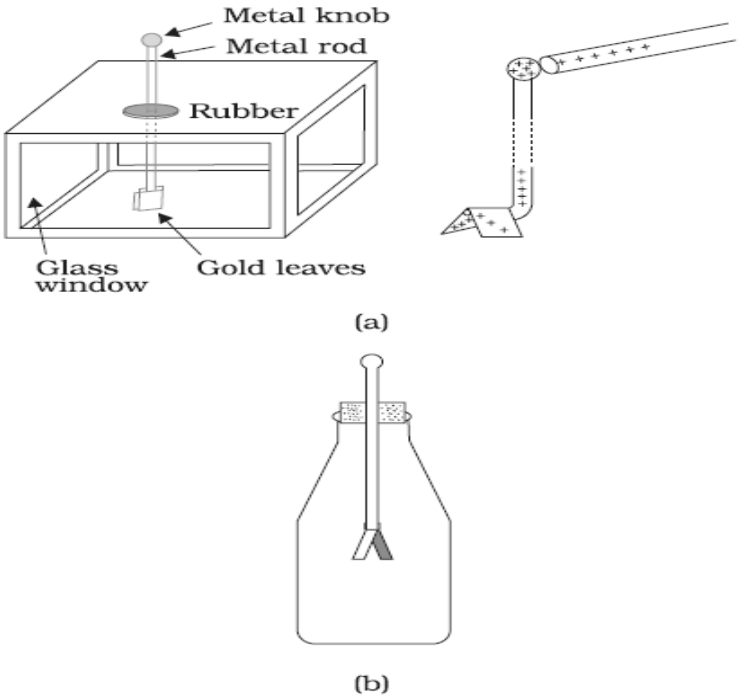


Fig. 4.2 Electroscopes: (a) The gold leaf electroscope, (b) Schematics of a simple electroscope

Students can make a simple electroscope as follows [Fig. 4.2(b)]: Take a thin aluminium curtain rod with ball ends fitted for hanging the curtain. Cut out a piece of length about 20 cm with the ball at one end and flatten the cut end. Take a large bottle that can hold this rod and a cork which will fit in the opening of the bottle. Make a hole in the cork sufficient to hold the curtain rod snugly. Slide the rod through the hole in the cork with the cut end on the lower side and ball end projecting above the cork. Fold a small, thin aluminium foil (about 6 cm in length) in the middle and attach it to the flattened end of the rod by cellulose tape. This forms the leaves of your electroscope. Fit the cork in the bottle with about 5 cm of the ball end projecting above the cork. A paper scale may be put inside the bottle in advance to measure the separation of leaves. The separation is a rough measure of the amount of charge on the electroscope.

To understand how the electroscope works, use the white paper strips we used for seeing the attraction of charged bodies. Fold the strips into half so that you make a mark of fold. Open the strip and iron it lightly with the mountain fold up, as shown in Fig. 4.3. Hold the strip by pinching it at the fold. You would notice that the two halves move apart.

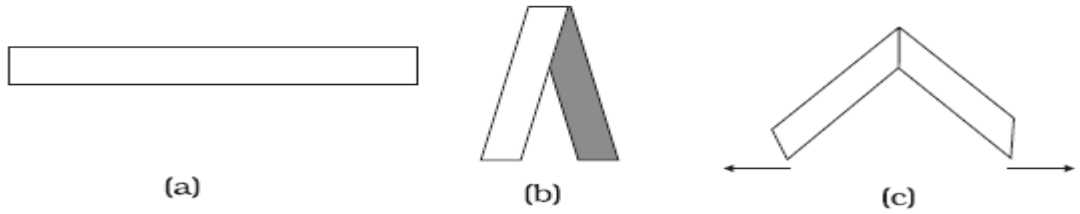


Fig. 1.3 Paper strip experiment.

This shows that the strip has acquired charge on ironing. When you fold it into half, both the halves have the same charge. Hence they repel each other. The same effect is seen in the leaf electroscope. On charging the curtain rod by touching the ball end with an electrified body, charge is transferred to the curtain rod and the attached aluminium foil. Both the halves of the foil get similar charge and therefore repel each other. The divergence in the leaves depends on the amount of charge on them. Let us first try to understand why material bodies acquire charge.

You know that all matter is made up of atoms and/or molecules. Although normally the materials are electrically neutral, they do contain charges; but their charges are exactly balanced. Forces that hold the molecules together, forces that hold atoms together in a solid, the adhesive force of glue, forces associated with surface tension, all are basically electrical in nature, arising from the forces between charged particles. Thus the electric force is all pervasive and it encompasses almost each and every field associated with our life. It is therefore essential that we learn more about such a force.

To electrify a neutral body, we need to add or remove one kind of charge. When we say that a body is charged, we always refer to this excess charge or deficit of charge. In solids, some of the electrons, being less tightly bound in the atom, are the charges which are transferred from one body to the other. A body can thus be charged positively by losing some of its electrons. Similarly, a body can be charged negatively by gaining electrons. When we rub a glass rod with silk, some of the electrons from the rod are transferred to the silk cloth. Thus the rod gets positively charged and the silk gets negatively charged. No new charge is created in the process of rubbing. Also the number of electrons, that are transferred, is a very small fraction of the total number of electrons in the material body. Also only the less tightly bound electrons in a material body can be transferred from it to another by rubbing. Therefore, when a body is rubbed with another, the bodies get charged and that is why we have to stick to certain pairs of materials to notice charging on rubbing the bodies.

4.3 CONDUCTORS AND INSULATORS

A metal rod held in hand and rubbed with wool will not show any sign of being charged. However, if a metal rod with a wooden or plastic handle is rubbed without touching its metal part, it shows signs of charging. Suppose we connect one end of a copper wire to a neutral pith ball and the other end to a negatively charged plastic rod. We will find that the pith ball acquires a negative charge. If a similar experiment is repeated with a nylon thread or a rubber band, no transfer of charge will take place from the plastic rod to the

pith ball. Why does the transfer of charge not take place from the rod to the ball?

Some substances readily allow passage of electricity through them, others do not. Those which allow electricity to pass through them easily are called conductors. They have electric charges (electrons) that are comparatively free to move inside the material. Metals, human and animal bodies and earth are conductors. Most of the non-metals like glass, porcelain, plastic, nylon, wood offer high resistance to the passage of electricity through them. They are called insulators. Most substances fall into one of the two classes stated above*.

When some charge is transferred to a conductor, it readily gets distributed over the entire surface of the conductor. In contrast, if some charge is put on an insulator, it stays at the same place. You will learn why this happens in the next chapter.

This property of the materials tells you why a nylon or plastic comb gets electrified on combing dry hair or on rubbing, but a metal article like spoon does not. The charges on metal leak through our body to the ground as both are conductors of electricity.

When we bring a charged body in contact with the earth, all the excess charge on the body disappears by causing a momentary current to pass to the ground through the connecting conductor (such as our body). This process of sharing the charges with the earth is called grounding or earthing. Earthing provides a safety measure for electrical circuits and appliances. A thick metal plate is buried deep into the earth and thick wires are drawn from this plate; these are used in buildings for the purpose of earthing near the mains supply. The electric wiring in our houses has three wires: live, neutral and earth. The first two carry electric current from the power station and the third is earthed by connecting it to the buried metal plate. Metallic bodies of the electric appliances such as electric iron, refrigerator, TV are connected to the earth wire. When any fault occurs or live wire touches the metallic body, the charge flows to the earth without damaging the appliance and without causing any injury to the humans; this would have otherwise been unavoidable since the human body is a conductor of electricity.

4.4 CHARGING BY INDUCTION

When we touch a pith ball with an electrified plastic rod, some of the negative charges on the rod are transferred to the pith ball and it also gets charged. Thus the pith ball is charged by contact. It is then repelled by the plastic rod but is attracted by a glass rod which is oppositely charged. However, why a electrified rod attracts light objects, is a question we have still left unanswered. Let us try to understand what could be happening by performing the following experiment.

---

* There is a third category called semiconductors, which offer resistance to the movement of charges which is intermediate between the conductors and insulators.

(i) Bring two metal spheres, A and B, supported on insulating stands, in contact as shown in Fig. 4.4(a).

(ii) Bring a positively charged rod near one of the spheres, say A, taking care that it does not touch the sphere. The free electrons in the spheres are attracted towards the rod. This leaves an excess of positive charge on the rear surface of sphere B. Both kinds of charges are bound in the metal spheres and cannot escape. They, therefore, reside on the surfaces, as shown in Fig. 4.4(b). The left surface of sphere A, has an excess of negative charge and the right surface of sphere B, has an excess of positive charge. However, not all of the electrons in the spheres have accumulated on the left surface of A. As the negative charge starts building up at the left surface of A, other electrons are repelled by these. In a short time, equilibrium is reached under the action of force of attraction of the rod and the force of repulsion due to the accumulated charges. Fig. 4.4(b) shows the equilibrium situation. The process is called induction of charge and happens almost instantly. The accumulated charges remain on the surface, as shown, till the glass rod is held near the sphere. If the rod is removed, the charges are not acted by any outside force and they redistribute to their original neutral state.

(iii) Separate the spheres by a small distance while the glass rod is still held near sphere A, as shown in Fig. 4.4(c). The two spheres are found to be oppositely charged and attract each other.

(iv) Remove the rod. The charges on spheres rearrange themselves as shown in Fig. 4.4(d). Now, separate the spheres quite apart. The charges on them get uniformly distributed over them, as shown in Fig. 4.4(e).

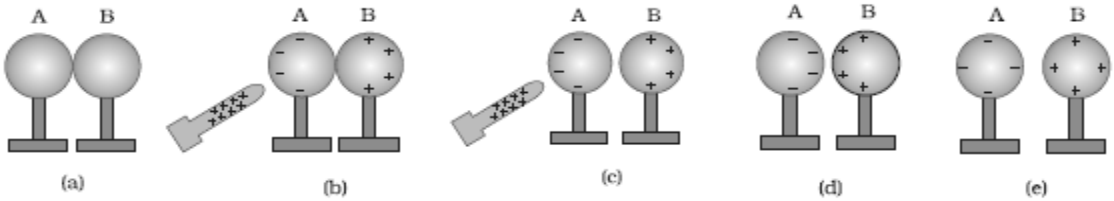


Fig. 4.4 Charging by induction.

In this process, the metal spheres will each be equal and oppositely charged. This is charging by induction. The positively charged glass rod does not lose any of its charge, contrary to the process of charging by contact.

When electrified rods are brought near light objects, a similar effect takes place. The rods induce opposite charges on the near surfaces of the objects and similar charges move to the farther side of the object. [This happens even when the light object is not a conductor. The mechanism for how this happens is explained later in Sections 4.10 and 5.10].

The centres of the two types of charges are slightly separated. We know that opposite charges attract while similar charges repel. However, the magnitude of force depends on the distance between the charges and in this case the force of attraction overweighs the force of repulsion. As a result the particles like bits of paper or pith balls, being light, are pulled towards the rods.

4.5 BASIC PROPERTIES OF ELECTRIC CHARGE

We have seen that there are two types of charges, namely positive and negative and their effects tend to cancel each other. Here, we shall now describe some other properties of the electric charge.

If the sizes of charged bodies are very small as compared to the distances between them, we treat them as point charges. All the charge content of the body is assumed to be concentrated at one point in space.

4.5.1 Additivity of charges

We have not as yet given a quantitative definition of a charge; we shall follow it up in the next section. We shall tentatively assume that this can be done and proceed. If a system contains two point charges q_1 and q_2, the total charge of the system is obtained simply by adding algebraically q_1 and q_2 , i.e., charges add up like real numbers or they are scalars like the mass of a body. If a system contains n charges q_1, q_2, q_3, …, q_n, then the total charge of the system is $q_1 + q_2 + q_3 + \ldots + q_n$. Charge has magnitude but no direction, similar to the mass. However, there is one difference between mass and charge. Mass of a body is always positive whereas a charge can be either positive or negative. Proper signs have to be used while adding the charges in a system. For example, the total charge of a system containing five charges +1, +2, –3, +4 and –5, in some arbitrary unit, is (+1) + (+2) + (–3) + (+4) + (–5) = –1 in the same unit.

4.5.2 Charge is conserved

We have already hinted to the fact that when bodies are charged by rubbing, there is transfer of electrons from one body to the other; no new charges are either created or destroyed. A picture of particles of electric charge enables us to understand the idea of conservation of charge. When we rub two bodies, what one body gains in charge the other body loses. Within an isolated system consisting of many charged bodies, due to interactions among the bodies, charges may get redistributed but it is found that the *total charge of the isolated system is always conserved.* Conservation of charge has been established experimentally.

It is not possible to create or destroy net charge carried by any isolated system although the charge carrying particles may be created or destroyed in a process. Sometimes nature creates charged particles: a neutron turns into a proton and an electron. The proton and electron thus created have equal and opposite charges and the total charge is zero before and after the creation.

4.5.3 Quantisation of charge

Experimentally it is established that all free charges are integral multiples of a basic unit of charge denoted by e. Thus charge q on a body is always given by

$$q = ne$$

where n is any integer, positive or negative. This basic unit of charge is the charge that an electron or proton carries. By convention, the charge on an electron is taken to be negative; therefore charge on an electron is written as –e and that on a proton as +e.

The fact that electric charge is always an integral multiple of e is termed as quantisation of charge. There are a large number of situations in physics where certain physical quantities are quantised. The quantisation of

charge was first suggested by the experimental laws of electrolysis discovered by English experimentalist Faraday. It was experimentally demonstrated by Millikan in 1912.

In the International System (SI) of Units, a unit of charge is called a coulomb and is denoted by the symbol C. A coulomb is defined in terms the unit of the electric current which you are going to learn in a subsequent chapter. In terms of this definition, one coulomb is the charge flowing through a wire in 1 s if the current is 1 A (ampere). In this system, the value of the basic unit of charge is

$$e = 1.602192 \times 10^{-19}\ C$$

Thus, there are about 6 x 10^{18} electrons in a charge of –1C. In electrostatics, charges of this large magnitude are seldom encountered and hence we use smaller units 1 μC (micro coulomb) = 10^{-6} C or 1 mC (milli coulomb) = 10^{-3} C.

If the protons and electrons are the only basic charges in the universe, all the observable charges have to be integral multiples of e. Thus, if a body contains n1 electrons and n2 protons, the total amount of charge on the body is $n_2\, x\, e + n_1\, x\, (-e) = (n_2 - n_1)e$. Since n_1 and n_2 are integers, their difference is also an integer. Thus the charge on any body is always an integral multiple of e and can be increased or decreased also in steps of e.

The step size e is, however, very small because at the macroscopic level, we deal with charges of a few μC. At this scale the fact that charge of a body can increase or decrease in units of e is not visible. The grainy nature of the charge is lost and it appears to be continuous.

This situation can be compared with the geometrical concepts of points and lines. A dotted line viewed from a distance appears continuous to us but is not continuous in reality. As many points very close to each other normally give an impression of a continuous line, many small charges taken together appear as a continuous charge distribution.

At the macroscopic level, one deals with charges that are enormous compared to the magnitude of charge e. Since e = 1.6 x 10^{-19} C, a charge of magnitude, say 1 μC, contains something like 10^{13} times the electronic charge. At this scale, the fact that charge can increase or decrease only in units of e is not very different from saying that charge can take continuous values. Thus, at the macroscopic level, the quantisation of charge has no practical consequence and can be ignored. At the microscopic level, where the charges involved are of the order of a few tens or hundreds of e, i.e., they can be counted, they appear in discrete lumps and quantisation of charge cannot be ignored. It is the scale involved that is very important.

4.6 COULOMB'S LAW

Coulomb's law is a quantitative statement about the force between two point charges. When the linear size of charged bodies are much smaller than the distance separating them, the size may be ignored and the charged bodies are treated as point charges. Coulomb measured the force between two point charges and found that it varied inversely as the square of the distance between the charges and was directly proportional to the product of the magnitude of the two charges and acted along the line joining the two charges. Thus, if two point charges q1, q2 are separated by a distance r in

vacuum, the magnitude of the force (F) between them is given by

$$F = k \frac{|q_1 q_2|}{r^2} \quad (4.1)$$

How did Coulomb arrive at this law from his experiments? Coulomb used a torsion balance* for measuring the force between two charged metallic spheres. When the separation between two spheres is much larger than the radius of each sphere, the charged spheres may be regarded as point charges. However, the charges on the spheres were unknown, to begin with. How then could he discover a relation like Eq. (4.1)? Coulomb thought of the following simple way: Suppose the charge on a metallic sphere is q. If the sphere is put in contact with an identical uncharged sphere, the charge will spread over the two spheres.

By symmetry, the charge on each sphere will be q/2*. Repeating this process, we can get charges q/2, q/4, etc. Coulomb varied the distance for a fixed pair of charges and measured the force for different separations. He then varied the charges in pairs, keeping the distance fixed for each pair. Comparing forces for different pairs of charges at different distances, Coulomb arrived at the relation, Eq. (4.1).

Coulomb's law, a simple mathematical statement, was initially experimentally arrived at in the manner described above. While the original experiments established it at a macroscopic scale, it has also been established down to subatomic level ($r \sim 10^{-10}$ m).

Charles Augustin de Coulomb (1736 – 1806) Coulomb, a French physicist, began his career as a military engineer in the West Indies. In 1776, he returned to Paris and retired to a small estate to do his scientific research. He invented a torsion balance to measure the quantity of a force and used it for determination of forces of electric attraction or repulsion between small charged spheres. He thus arrived in 1785 at the inverse square law relation, now known as Coulomb's law. The law had been anticipated by Priestley and also by

---

* A torsion balance is a sensitive device to measure force. It was also used later by Cavendish to measure the very feeble gravitational force between two objects, to verify Newton's Law of Gravitation.

---

* Implicit in this is the assumption of additivity of charges and conservation: two charges (*q/2* each) add up to make a total charge *q*.

Coulomb discovered his law without knowing the explicit magnitude of the charge. In fact, it is the other way round: Coulomb's law can now be employed to furnish a definition for a unit of charge. In the relation, Eq. (4.1), k is so far arbitrary. We can choose any positive value of k. The choice of k determines the size of the unit of charge. In SI units, the value of k is about 9×10^9. The unit of charge that results from this choice is called a coulomb which we defined earlier in Section 1.4. Putting this value of k in Eq. (4.1), we see that for $q_1 = q_2 = 1$ C, r = 1 m

$F = 9 \times 10^9$ N

That is, 1C is the charge that when placed at a distance of 1 m from another charge of the same magnitude in vacuum experiences an electrical force of repulsion of magnitude 9×10^9 N. One coulomb is evidently too big a unit to be used. In practice, in electrostatics, one uses smaller units like 1 mC or 1 μC.

The constant k in Eq. (4.1) is usually put as $k = 1/4\pi\epsilon_0$ for later convenience, so that Coulomb's law is written as

$$F = \frac{1}{4\pi\epsilon_0}\frac{|q_1 q_2|}{r^2} \qquad (4.2)$$

ϵ_0 is called the permittivity of free space. The value of $\epsilon 0$ in SI units is $\varepsilon_0 = 8.854 \times 10^{-12}\ C^2\ N^{-1}\ m^{-2}$.

Since force is a vector, it is better to write Coulomb's law in the vector notation. Let the position vectors of charges q_1 and q_2 be $\mathbf{r_1}$ and $\mathbf{r_2}$ respectively [see Fig. 4.5(a)]. We denote force on q_1 due to q_2 by $\mathbf{F}_{12}$ and force on q_2 due to q_1 by $\mathbf{F}_{21}$. The two point charges q_1 and q_2 have been numbered 1 and 2 for convenience and the vector leading from 1 to 2 is denoted by $\mathbf{r}_{21}$:

$$\boldsymbol{r_{21}} = \boldsymbol{r_2} - \boldsymbol{r_1}$$

The magnitude of the vectors $\mathbf{r}_{21}$ and $\mathbf{r}_{12}$ is denoted by $\mathbf{r}_{21}$ and $\mathbf{r}_{12}$, respectively ($r_{12} = r_{21}$). The direction of a vector is specified by a unit vector along the vector. To denote the direction from 1 to 2 (or from 2 to 1), we define the unit vectors:

$$\hat{\boldsymbol{r}}_{21} = \frac{\boldsymbol{r}_{21}}{r_{21}}, \qquad \hat{\boldsymbol{r}}_{12} = \frac{\boldsymbol{r}_{12}}{r_{12}}, \qquad \hat{\boldsymbol{r}}_{21} = \hat{\boldsymbol{r}}_{12}$$

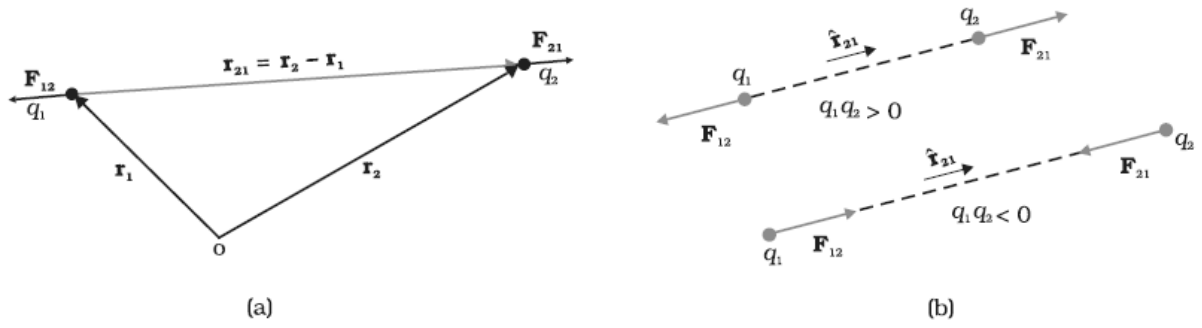


Fig. 4.5 (a) Geometry and (b) Forces between charges.

Coulomb's force law between two point charges q_1 and q_2 located at r_1 and r_2 is then expressed as

$$F_{21} = \frac{1}{4\pi\varepsilon_0} \frac{q_1 q_2}{r_{21}^2} \hat{r}_{21} \quad (4.3)$$

Some remarks on Eq. 4.3 are relevant:

- Equation (4.3) is valid for any sign of q_1 and q_2 whether positive or negative. If q_1 and q_2 are of the same sign (either both positive or both negative), $\mathbf{F}_{21}$ is along $\hat{\mathbf{r}}_{21}$, which denotes repulsion, as it should be for like charges. If q_1 and q_2 are of opposite signs, $\mathbf{F}_{21}$is along $-\hat{\mathbf{r}}_{21} (= \hat{\mathbf{r}}_{12})$, which denotes attraction, as expected for unlike charges. Thus, we do not have to write separate equations for the cases of like and unlike charges. Equation (4.3) takes care of both cases correctly [Fig. 4.5(b)].
- The force F_{12} on charge q1 due to charge q2, is obtained from Eq. (4.3), by simply interchanging 1 and 2, i.e.,

$$F_{12} = \frac{1}{4\pi\varepsilon_0} \frac{q_1 q_2}{r_{12}^2} \hat{r}_{12} = -F_{21}$$

Thus, Coulomb's law agrees with the Newton's law.

- Coulomb's law [Eq. (4.3)] gives the force between two charges q1 and q2 in vacuum. If the charges are placed in matter or the intervening space has matter, the situation gets complicated due to the presence of charged constituents of matter. We shall consider electrostatics in matter in the next chapter.

4.7 FORCES BETWEEN MULTIPLE CHARGES

The mutual electric force between two charges is given by Coulomb's law. How to calculate the force on a charge where there are not one but several charges around? Consider a system of *n* stationary charges q_1, q_2, q_3, ..., q_n in vacuum. What is the force on q_1 due to q_2, q_3, ..., q_n? Coulomb's law is not enough to answer this question. Recall that forces of mechanical origin add according to the parallelogram law of addition. Is the same true for forces of electrostatic origin?

Experimentally it is verified that *force on any charge due to a number of other charges is the vector sum of all the forces on that charge due to the other charges, taken one at a time. The individual forces are unaffected due to the presence of other charges*. This is termed as the *principle of superposition***.**

To better understand the concept, consider a system of three charges q_1, q_2 and q_3, as shown in Fig. 4.6(a).

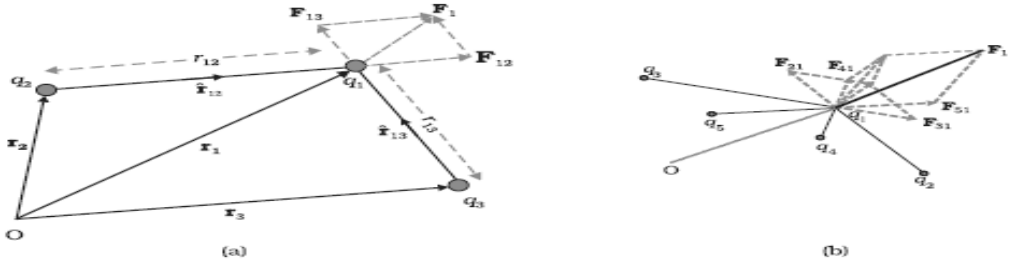


Fig. 4.6 A system of (a) three charges (b) multiple charges.

The force on one charge, say q_1, due to two other charges q_2, q_3 can therefore be obtained by performing a vector addition of the forces due to each one of these charges. Thus, if the force on q_1 due to q_2 is denoted by $\boldsymbol{F_{12}}$, $\boldsymbol{F_{12}}$ is given by Eq. (4.3) even though other charges are present.

Thus, $$F_{12} = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r_{12}^2}\hat{r}_{12}$$

In the same way, the force on q_1 due to q_3, denoted by F_{13}, is given by

$$F_{13} = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_3}{r_{13}^2}\hat{r}_{13}$$

which again is the Coulomb force on q_1 due to q_3, even though other charge q_2 is present.

Thus the total force F_1 on q_1 due to the two charges q_2 and q_3 is given as

$$\mathrm{F_1 = F_{12} + F_{13}} = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r_{12}^2}\hat{r}_{12} + \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_3}{r_{13}^2}\hat{r}_{13} \qquad (4.4)$$

The above calculation of force can be generalized to a system of charges more than three, as shown in Fig. 4.6(b).

The principle of superposition says that in a system of charges q_1, q_2, ..., q_n, the force on q_1 due to q_2 is the same as given by Coulomb's law, i.e., it is unaffected by the presence of the other charges q_3, q_4, ..., q_n. The total force F_1 on the charge q_1, due to all other charges, is then given by the vector sum of the forces $F_{12}, F_{13}, \ldots, F_{1n}$:

$$F_1 = F_{12} + F_{13} + \ldots \mathrm{F_{1n}} = \frac{1}{4\pi\varepsilon_0}\left[\frac{q_1 q_2}{r_{12}^2}\hat{r}_{12} + \frac{q_1 q_3}{r_{13}^2}\hat{r}_{13} + \frac{q_1 q_n}{r_{1n}^2}\hat{r}_{1n}\right]$$

$$= \frac{q_1}{4\pi\varepsilon_0}\sum_{i=2}^{n}\frac{q_i}{r_{1i}^2}\hat{r}_{1i} \qquad (4.5)$$

The vector sum is obtained as usual by the parallelogram law of addition of vectors. All of electrostatics is basically a consequence of Coulomb's law and the superposition principle.

4.8 ELECTRIC FIELD

Let us consider a point charge Q placed in vacuum, at the origin O. If we place another point charge q at a point P, where **OP** = **r**, then the charge Q will exert a force on q as per Coulomb's law. We may ask the question: If charge q is removed, then what is left in the surrounding? Is there nothing? If there is nothing at the point P, then how does a force act when we place the charge q at P.

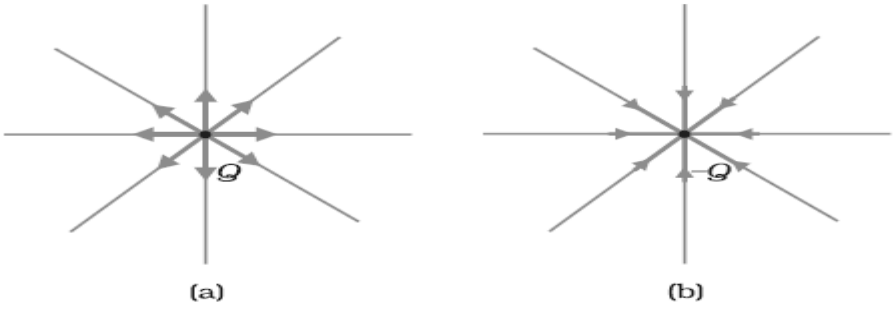


Fig. 4.7 Electric field (a) due to a charge Q, (b) due to a charge -Q

In order to answer such questions, the early scientists introduced the concept of field. According to this, we say that the charge Q produces an electric field everywhere in the surrounding. When another charge q is brought at some point P, the field there acts on it and produces a force. The electric field produced by the charge Q at a point r is given as

$$E(r) = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\hat{r} = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\hat{r} \quad (4.6)$$

where $\hat{\mathrm{r}} = \frac{\mathrm{r}}{\mathrm{r}}$, , is a unit vector from the origin to the point r. Thus, Eq.(4.6) specifies the value of the electric field for each value of the position vector r The word "field" signifies how some distributed quantity (which could be a scalar or a vector) varies with position. The effect of the charge has been incorporated in the existence of the electric field. We obtain the force F exerted by a charge Q on a charge q, as

$$F = \frac{1}{4\pi\varepsilon_0}\frac{Qq}{r^2}\hat{r} \quad (4.7)$$

Note that the charge q also exerts an equal and opposite force on the charge Q. The electrostatic force between the charges Q and q can be looked upon as an interaction between charge q and the electric field of Q and vice versa. If we denote the position of charge q by the vector r, it experiences a force F equal to the charge q multiplied by the electric field E at the location of q. Thus,

$$F(r) = q\,E(r) \quad (4.8)$$

Equation (4.8) defines the SI unit of electric field as N/C*.

(i) From Eq. (4.8) we can infer that if q is unity, the electric field due to a charge Q is numerically equal to the force exerted by it. Thus, the *electric field due to a charge Q at a point in space may be defined as the force that a unit positive charge would experience if placed at that point*. The charge Q, which is producing the electric field, is called a *source charge* and the charge q, which tests the effect of a source charge, is called a *test charge*. Note that the source charge Q must remain at its original location. However, if a charge q is brought at any point around Q, Q itself is bound to experience an electrical force due to q and will tend to move. A way out of this difficulty is to make q negligibly small. The force **F** is then negligibly small but the ratio $\mathbf{F}/q$ is finite and defines the electric field:

$$E = \lim_{q\to 0}\frac{F}{q} \quad (4.9)$$

A practical way to get around the problem (of keeping Q undisturbed in the presence of q) is to hold Q to its location by unspecified forces! This may look strange but actually this is what happens in practice. When we are considering the electric force on a test charge q due to a charged planar sheet (Section 4.15), the charges on the sheet are held to their locations by the forces due to the unspecified charged constituents inside the sheet.

---

* An alternate unit V/m will be introduced in the next chapter

(ii) Note that the electric field **E** due to Q, though defined operationally in terms of some test charge q, is independent of q. This is because **F** is proportional to q, so the ratio **F**/q does not depend on q. The force **F** on the charge q due to the charge Q depends on the particular location of charge q which may take any value in the space around the charge Q. Thus, the electric field **E** due to Q is also dependent on the space coordinate **r**. For different positions of the charge q all over the space, we get different values of electric field **E**. The field exists at every point in three-dimensional space.

(iii) For a positive charge, the electric field will be directed radially outwards from the charge. On the other hand, if the source charge is negative, the electric field vector, at each point, points radially inwards.

(iv) Since the magnitude of the force **F** on charge q due to charge Q depends only on the distance **r** of the charge q from charge Q, the magnitude of the electric field **E** will also depend only on the distance **r**. Thus at equal distances from the charge Q, the magnitude of its electric field **E** is same. The magnitude of electric field **E** due to a point charge is thus same on a sphere with the point charge at its centre; in other words, it has a spherical symmetry.

4.8.1 Electric field due to a system of charges

Consider a system of charges q_1, q_2, ..., q_n with position vectors $\mathbf{r_1}$, $\mathbf{r_2}$, ..., $\mathbf{r_n}$ relative to some origin O. Like the electric field at a point in space due to a single charge, electric field at a point in space due to the system of charges is defined to be the force experienced by a unit test charge placed at that point, without disturbing the original positions of charges q_1, q_2, ..., q_n. We can use Coulomb's law and the superposition principle to determine this field at a point P denoted by position vector **r**.

Electric field E1 at **r** due to q_1 at $\mathbf{r_1}$ is given by

$$E_1 = \frac{1}{4\pi\varepsilon_0}\frac{q_1}{r_{1P}^2}\hat{r}_{1P}$$

where $\hat{\mathbf{r}}_{\mathbf{1P}}$ is a unit vector in the direction from q_1 to P, and $\mathbf{r_{1P}}$ is the distance between q_1 and P.

In the same manner, electric field $\mathbf{E_2}$ at r due to q_2 at $\mathbf{r_2}$ is

$$E_2 = \frac{1}{4\pi\varepsilon_0}\frac{q_2}{r_{2P}^2}\hat{r}_{2P}$$

where $\hat{\mathbf{r}}_{\mathbf{2P}}$ is a unit vector in the direction from q_2 to P, and $\mathbf{r_{2P}}$ is the distance between q_1 and P. Similar expressions hold good for fields E_3, E_4, …, E_n due to charges q_3, q_4,…,q_n.

By the superposition principle, the electric field **E** at **r** due to the system of charges is (as shown in Fig. 4.8)

$$\mathbf{E(r) = E_1 (r) + E_2 (r) + \ldots + E_n (r)}$$

$$= \frac{1}{4\pi\varepsilon_0}\frac{q_1}{r_{1P}^2}\hat{r}_{1P} + \frac{1}{4\pi\varepsilon_0}\frac{q_2}{r_{2P}^2}\hat{r}_{2P} + \cdots + \frac{1}{4\pi\varepsilon_0}\frac{q_n}{r_{nP}^2}\hat{r}_{nP}$$

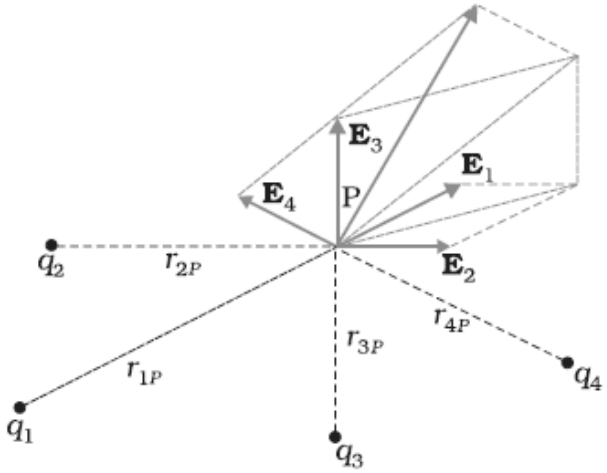


Fig. 4.8 Electric field at a point due to a system of charges is the vector sum of the electric fields at the point due to individual charges.

$$\mathbf{E(r)} = \frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{n}\frac{q_i}{r_{ip}^2}\hat{r}_{ip} \qquad (4.10)$$

E is a vector quantity that varies from one point to another point in space and is determined from the positions of the source charges.

4.8.2 Physical significance of electric field

You may wonder why the notion of electric field has been introduced here at all. After all, for any system of charges, the measurable quantity is the force on a charge which can be directly determined using Coulomb's law and the superposition principle [Eq. (4.5)]. Why then introduce this intermediate quantity called the electric field?

For electrostatics, the concept of electric field is convenient, but not really necessary. Electric field is an elegant way of characterising the electrical environment of a system of charges. Electric field at a point in the space around a system of charges tells you the force a unit positive test charge would experience if placed at that point (without disturbing the system). Electric field is a characteristic of the system of charges and is independent of the test charge that you place at a point to determine the field. The term *field* in physics generally refers to a quantity that is defined at every point in space and may vary from point to point. Electric field is a vector field, since force is a vector quantity.

The true physical significance of the concept of electric field, however, emerges only when we go beyond electrostatics and deal with time-dependent electromagnetic phenomena. Suppose we consider the force between two distant charges q_1, q_2 in accelerated motion. Now the greatest speed with which a signal or information can go from one point to another is c, the speed of light. Thus, the effect of any motion of q_1 on q_2 cannot arise instantaneously. There will be some time delay between the effect (force on q_2) and the cause (motion of q_1). It is precisely here that the notion of electric field (strictly, electromagnetic field) is natural and very useful. *The field picture is this: the accelerated motion of charge q_1 produces electromagnetic waves, which then propagate with the speed c, reach q_2 and cause a force on q_2.* The notion of field elegantly accounts for the time delay. Thus, even

though electric and magnetic fields can be detected only by their effects (forces) on charges, they are regarded as physical entities, not merely mathematical constructs. They have an *independent dynamics* of their o*wn*, i.e., they evolve according to laws of their *own*. They can also transport energy. Thus, a source of time- dependent electromagnetic fields, turned on briefly and switched off, leaves behind propagating electromagnetic fields transporting energy. The concept of field was first introduced by Faraday and is now among the central concepts in physics.

4.9 ELECTRIC FIELD LINES

We have studied electric field in the last section. It is a vector quantity and can be represented as we represent vectors. Let us try to represent E due to a point charge pictorially. Let the point charge be placed at the origin. Draw vectors pointing along the direction of the electric field with their lengths proportional to the strength of the field at each point. Since the magnitude of electric field at a point decreases inversely as the square of the distance of that point from the charge, the vector gets shorter as one goes away from the origin, always pointing radially outward. Figure 4.9 shows such a picture.

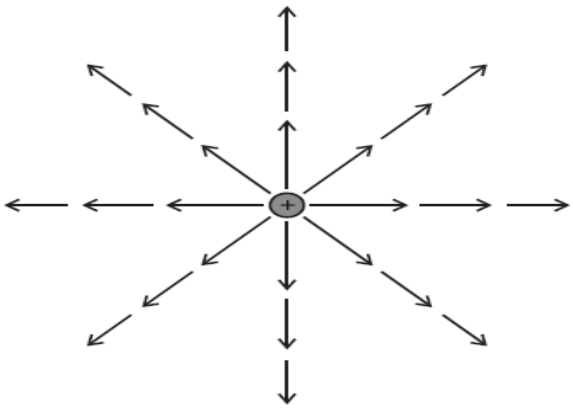


Fig. 4.9 Field of a point charge

In this figure, each arrow indicates the electric field, i.e., the force acting on a unit positive charge, placed at the tail of that arrow. Connect the arrows pointing in one direction and the resulting figure represents a field line. We thus get many field lines, all pointing outwards from the point charge. Have we lost the information about the strength or magnitude of the field now, because it was contained in the length of the arrow? No. Now the magnitude of the field is represented by the density of field lines. **E** is strong near the charge, so the density of field lines is more near the charge and the lines are closer. Away from the charge, the field gets weaker and the density of field lines is less, resulting in well-separated lines.

Another person may draw more lines. But the number of lines is not important. In fact, an infinite number of lines can be drawn in any region. It is the relative density of lines in different regions which is important.

We draw the figure on the plane of paper, i.e., in two- dimensions but we live in three-dimensions. So if one wishes to estimate the density of field lines, one has to consider the number of lines per unit cross-sectional area, perpendicular to the lines. Since the electric field decreases as the square of

the distance from a point charge and the area enclosing the charge increases as the square of the distance, the number of field lines crossing the enclosing area remains constant, whatever may be the distance of the area from the charge.

We started by saying that the field lines carry information about the direction of electric field at different points in space. Having drawn a certain set of field lines, the relative density (i.e., closeness) of the field lines at different points indicates the relative strength of electric field at those points. The field lines crowd where the field is strong and are spaced apart where it is weak. Figure 4.10 shows a set of field lines. We can imagine two equal and small elements of area placed at points R and S normal to the field lines there. The number of field lines in our picture cutting the area elements is proportional to the magnitude of field at these points. The picture shows that the field at R is stronger than at S.

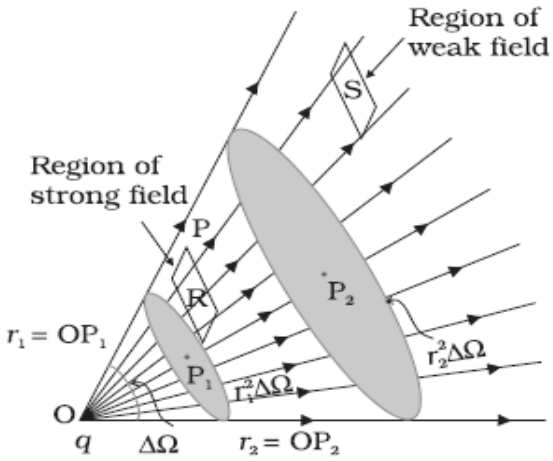


Fig. 1.16 Dependence of electric field strength on the distance and its relation to the number of field lines.

To understand the dependence of the field lines on the area, or rather the *solid angle* subtended by an area element, let us try to relate the area with the solid angle, a generalization of angle to three dimensions. Recall how a (plane) angle is defined in two-dimensions. Let a small transverse line element Δl be placed at a distance r from a point O. Then the angle subtended by Δl be placed at a distance r from a point O. Then the angle subtended by Δl at O can be approximated as $\Delta\theta = \Delta l/r$. Likewise, in three-dimensions the solid angle $*$ subtended by a small perpendicular plane area ΔS, at a distance r, can be written as $\Delta W = \Delta s/r^2$. We know that in a given solid angle the number of radial field lines is the same. In Fig 4.10 for two points P_1 and P_2 at distances r_1 and r_2 from the charge, the element of area subtending the solid angle ΔW is $r_1^2\Delta W$ at P_1 and an element of area $r_2^2\Delta W$ at P_2, respectively. The number of lines (say n) cutting unit area element is therefore $n/(r_1^2\Delta W)$ at P_1 and $n/(r_2^2\Delta W)$ at P_2, respectively. Since n and ΔW are common, the strength of the field clearly has a $1/r^2$ dependence.

---

* Solid angle is a measure of a cone. Consider the intersection of the given cone with a sphere of radius R. The solid angle $\Delta\Omega$ of the cone is defined to be equal to $\Delta S/R^2$, where ΔS is the area on the sphere cut out by the cone

The picture of field lines was invented by Faraday to develop an intuitive non- mathematical way of visualizing electric fields around charged configurations. Faraday called them *lines of force*. *This term is somewhat misleading, especially in case of magnetic fields. The more appropriate term is field lines (electric or magnetic) that we have adopted in this book.*

Electric field lines are thus a way of pictorially mapping the electric field around a configuration of charges. An electric field line is, in general a curve drawn in such a way that the tangent to it at each point is in the direction of the net field at that point. An arrow on the curve is obviously necessary to specify the direction of electric field from the two possible directions indicated by a tangent to the curve. A field line is a space curve, i.e., a curve in three dimensions.

Figure 4.11 shows the field lines around some simple charge configurations. As mentioned earlier, the field lines are in 3-dimensional space, though the figure shows them only in a plane. The field lines of a single positive charge are radially outward while those of a single negative charge are radially inward. The field lines around a system of two positive charges *(q, q)* give a vivid pictorial description of their mutual repulsion, while those around the configuration of two equal and opposite charges *(q, –q)*, a dipole, show clearly the mutual attraction between the charges. The field lines follow some important general properties:

(i) Field lines start from positive charges and end at negative charges. If there is a single charge, they may start or end at infinity.

(ii) In a charge-free region, electric field lines can be taken to be continuous curves without any breaks.

(iii) Two field lines can never cross each other. (If they did, the field at the point of intersection will not have a unique direction, which is absurd.)

(iv) Electrostatic field lines do not form any closed loops. This follows from the conservative nature of electric field (Chapter 5).

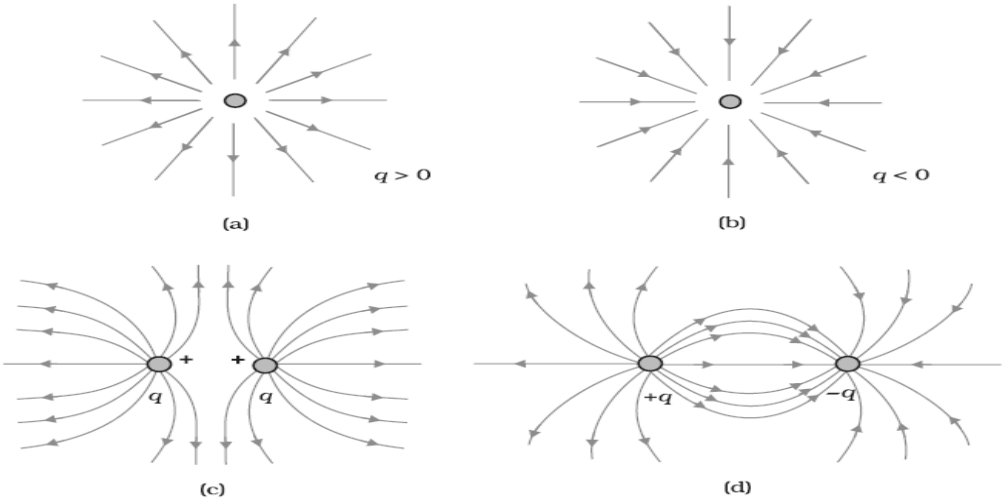


Fig. 4.11 Field lines due to some simple charge configurations

4.10 ELECTRIC FLUX

Consider flow of a liquid with velocity **v**, through a small flat surface dS, in a direction normal to the surface. The rate of flow of liquid is given by the volume crossing the area per unit time *v*dS and represents the flux of liquid flowing across the plane. If the normal to the surface is not parallel to the direction of flow of liquid, i.e., to **v**, but makes an angle θ with it, the projected area in a plane perpendicular to **v** is v dS cos θ. Therefore the flux going out of the surface dS is $\mathbf{v}.\hat{\boldsymbol{n}}$ dS. For the case of the electric field, we define an analogous quantity and call it *electric flux*. We should however note that there is no *flow* of a physically observable quantity unlike the case of liquid flow.

In the picture of electric field lines described above, we saw that the number of field lines crossing a unit area, placed normal to the field at a point is a measure of the strength of electric field at that point. This means that if we place a small planar element of area ΔS normal to E at a point, the number of field lines crossing it is proportional* to E ΔS. Now suppose we tilt the area element by angle θ. Clearly, the number of field lines crossing the area element will be smaller. The projection of the area element normal to E is ΔS cosθ. Thus the number of field lines crossing ΔSis proportional to EΔS cosθ. When θ = 90°, field lines will be parallel to ΔSand will not cross it at all (Fig. 4.12).

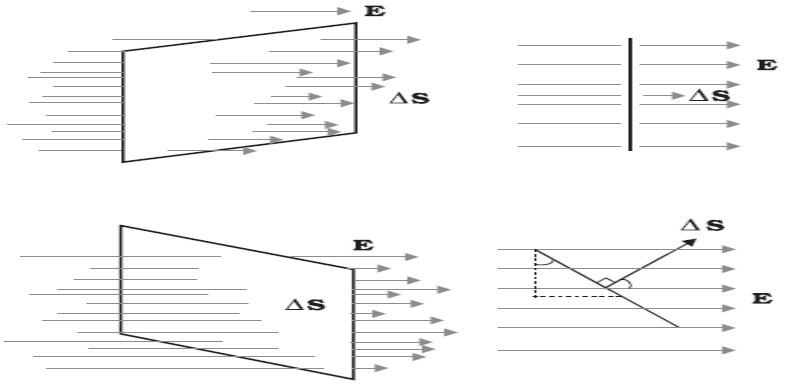


Fig. 4.12 Dependence of flux on the inclination θ between E and $\hat{n}$.

The orientation of area element and not merely its magnitude is important in many contexts. For example, in a stream, the amount of water flowing through a ring will naturally depend on how you hold the ring. If you hold it normal to the flow, maximum water will flow through it than if you hold it with some other orientation. This shows that an area element should be treated as a vector. It has a magnitude and also a direction. How to specify the direction of a planar area? Clearly, the normal to the plane specifies the orientation of the plane. Thus the direction of a planar area vector is along its normal.

---

* It will not be proper to say that the number of field lines is equal to EΔS. The number of field lines is after all, a matter of how many field lines we choose to draw. What is physically significant is the relative number of field lines crossing a given area at different points.

How to associate a vector to the area of a curved surface? We imagine dividing the surface into a large number of very small area elements. Each small area element may be treated as planar and a vector associated with it, as explained before.

Notice one ambiguity here. The direction of an area element is along its normal. But a normal can point in two directions. Which direction do we choose as the direction of the vector associated with the area element? This problem is resolved by some convention appropriate to the given context. For the case of a closed surface, this convention is very simple. The vector associated with every area element of a closed surface is taken to be in the direction of the *outward normal*. This is the convention used in Fig 4.13. Thus, the area element vector $\Delta\mathbf{S}$ at a point on a closed surface equals $\Delta S\,\hat{\mathbf{n}}$ where ΔS is the magnitude of the area element and $\hat{\mathbf{n}}$ is a vector in the direction of outward normal at that point.

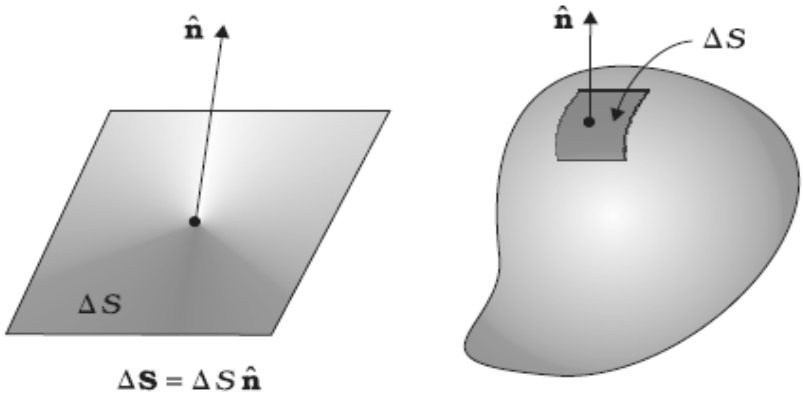


Fig. 4.13 Convention for defining normal $\hat{n}$ and ΔS.

We now come to the definition of electric flux. Electric flux $\Delta\emptyset$ through an area element $\Delta S$is defined by

$$\Delta\emptyset = \boldsymbol{E}.\Delta\boldsymbol{S} = E\Delta S\cos\theta \qquad (4.11)$$

which, as seen before, is proportional to the number of field lines cutting the area element. The angle θ here is the angle between **E** and $\Delta\mathbf{S}$. For a closed surface, with the convention stated already, θ is the angle between **E** and the outward normal to the area element. Notice we could look at the expression $E\Delta S\cos\theta$in two ways: $E(\Delta S\cos\theta)$i.e., E times the projection of area normal to **E**, or $E^{\perp}\,\Delta S$, i.e., component of **E** along the normal to the area element times the magnitude of the area element. The unit of electric flux is N C^{-1} m^2.

The basic definition of electric flux given by Eq. (4.11) can be used, in principle, to calculate the total flux through any given surface. All we have to do is to divide the surface into small area elements, calculate the flux at each element and add them up. Thus, the total flux $\emptyset$ through a surface S is

$$\emptyset \simeq \sum \boldsymbol{E}.\Delta\boldsymbol{S} \qquad (4.12)$$

The approximation sign is put because the electric field **E** is taken to be constant over the small area element. This is mathematically exact only when you take the limit $\Delta S \to 0$ and the sum in Eq. (4.12) is written as an integral.

4.11 ELECTRIC DIPOLE

An electric dipole is a pair of equal and opposite point charges q and $-q$,

separated by a distance 2a. The line connecting the two charges defines a direction in space. By convention, the direction from $-q$ to q is said to be the direction of the dipole. The mid-point of locations of $-q$ and q is called the centre of the dipole.

The total charge of the electric dipole is obviously zero. This does not mean that the field of the electric dipole is zero. Since the charge q and $-q$ are separated by some distance, the electric fields due to them, when added, do not exactly cancel out. However, at distances much larger than the separation of the two charges forming a dipole (r >> 2a), the fields due to q and $-q$ nearly cancel out. The electric field due to a dipole therefore falls off, at large distance, faster than like $1/r^2$ (the dependence on r of the field due to a single charge q). These qualitative ideas are borne out by the explicit calculation as follows:

4.11.1 The field of an electric dipole

The electric field of the pair of charges ($-q$ and q) at any point in space can be found out from Coulomb's law and the superposition principle. The results are simple for the following two cases: (i) when the point is on the dipole axis, and (ii) when it is in the equatorial plane of the dipole, i.e., on a plane perpendicular to the dipole axis through its centre. The electric field at any general point P is obtained by adding the electric fields E_{-q} due to the charge –q and E_{+q} due to the charge q, by the parallelogram law of vectors.

(i) For points on the axis

Let the point P be at distance r from the centre of the dipole on the side of the charge q, as shown in Fig. 4.14(a). Then

$$\mathbf{E}_{-q} = \frac{q}{4\pi\varepsilon_0(r+a)^2}\boldsymbol{P} \qquad (4.13(a))$$

where $\hat{p}$ is the unit vector along the dipole axis (from $-q$ to q). Also

$$\mathbf{E}_{+q} = \frac{q}{4\pi\varepsilon_0(r-a)^2}\boldsymbol{P} \qquad (4.13\ (b))$$

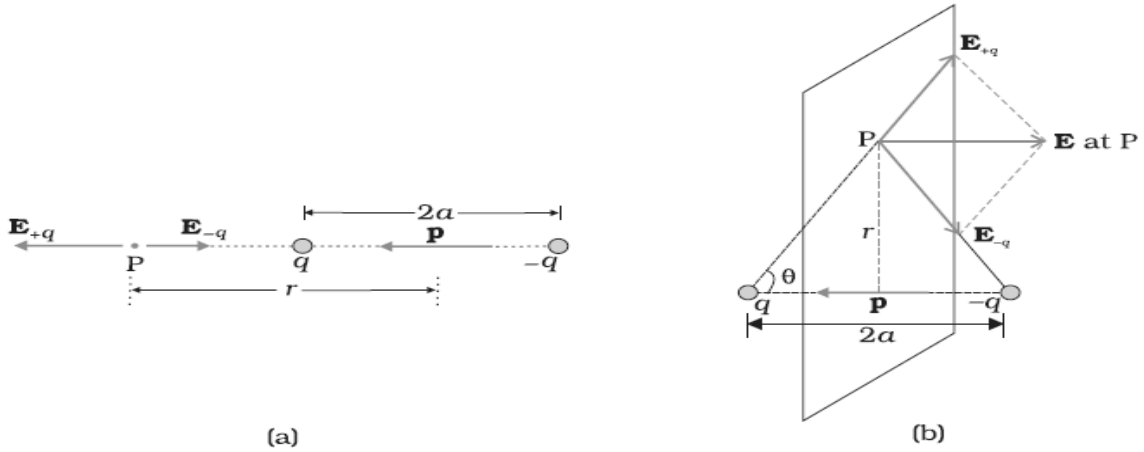


Fig. 1.20 Electric field of a dipole at (a) a point on the axis, (b) a point on the equatorial plane of the dipole. p is the dipole moment vector of magnitude $p = q \times 2a$ and directed from $-q$ *to* q.

The total field P is

$$E = E_{+q} + E_{-q} = \frac{q}{4\pi\epsilon_0}\left[\frac{1}{(r-a)^2} - \frac{1}{(r+a)^2}\right]\boldsymbol{p}$$

$$= \frac{q}{4\pi\varepsilon_0}\ \frac{4\,a\,r}{(r^2-q^2)^2}\ \boldsymbol{p} \qquad \textbf{(}4.14\textbf{)}$$

For r>>a

$$\mathbf{E} = \frac{4qa}{4\pi\varepsilon_0 r^3}\hat{\boldsymbol{p}} \qquad \text{(r>>a)} \qquad (4.15)$$

(ii) For points on the equatorial plane

The magnitudes of the electric fields due to the two charges $+q$ and $-q$ are given by

$$E_{+q} = \frac{q}{4\pi\varepsilon_0}\ \frac{1}{r^2+a^2} \qquad 4.16\ (a)$$

$$E_{-q} = \frac{q}{4\pi\varepsilon_0}\ \frac{1}{r^2-a^2} \qquad 4.16\ (b)$$

The directions of E_{+q} and E_{-q} are as shown in Fig. 4.14(b). Clearly, the components normal to the dipole axis cancel away. The components along the dipole axis add up. The total electric field is opposite to $\hat{p}$. We have

$$E = -(E_{+q} + E_{-q})\cos\theta\ \hat{\boldsymbol{p}}$$

$$= \frac{2qa}{4\pi\varepsilon_0\ (r^2+a^2)^{\frac{3}{2}}}\ \boldsymbol{p} \qquad \textbf{(}4.17\textbf{)}$$

At large distances (r>> a), this reduces to

$$\mathbf{E} = \frac{2\,q\,a}{4\pi\varepsilon_0 r^3}\ \hat{\boldsymbol{p}} \quad \text{(r>> } a\textbf{)} \qquad \textbf{(}4.18\textbf{)}$$

From Eqs. 4.15 and 4.18, it is clear that the dipole field at large distances does not involve q and a separately; it depends on the product qa. This suggests the definition of dipole moment. The *dipole moment vector* **p** of an electric dipole is defined by

$$\mathbf{p} = \text{q x } 2a\ \hat{\boldsymbol{p}} \qquad (4.19)$$

that is, it is a vector whose magnitude is charge q times the separation $2a$ (between the pair of charges $q, -q$) and the direction is along the line from $-q$ to q. In terms of **p**, the electric field of a dipole at large distances takes simple forms:

At a point on the dipole axis

$$\boldsymbol{E} = \frac{2\,\boldsymbol{p}}{4\pi\varepsilon_0 r^3} \qquad (r>>a) \qquad (4.20)$$

At a point on the equatorial plane

$$\boldsymbol{E} = -\frac{2\,\boldsymbol{p}}{4\pi\varepsilon_0 r^3} \qquad (r>>a) \qquad (4.21)$$

Notice the important point that the dipole field at large distances falls off not as $1/r^2$ but as$1/r^3$. Further, the magnitude and the direction of the dipole field depends not only on the distance r but also on the angle between the position vector r and the dipole moment **p.**

We can think of the limit when the dipole size 2a approaches zero, the charge q approaches infinity in such a way that the product $p = q \times 2a$ is finite. Such a dipole is referred to as a point dipole. For a point dipole, Eqs. (4.20) and (4.21) are exact, true for any r.

4.11.2 Physical significance of dipoles

In most molecules, the centres of positive charges and of negative charges* lie at the same place. Therefore, their dipole moment is zero. CO_2 and CH_4 are of this type of molecules. However, they develop a dipole moment when an electric field is applied. But in some molecules, the centres of negative charges and of positive charges do not coincide. Therefore they have a permanent electric dipole moment, even in the absence of an electric field. Such molecules are called polar molecules. Water molecules, H_2O, is an example of this type. Various materials give rise to interesting properties and important applications in the presence or absence of electric field.

4.12 DIPOLE IN A UNIFORM EXTERNAL FIELD

Consider a permanent dipole of dipole moment **p** in a uniform external field **E**, as shown in Fig. 4.15. (By permanent dipole, we mean that **p** exists irrespective of **E**; it has not been induced by **E**.)

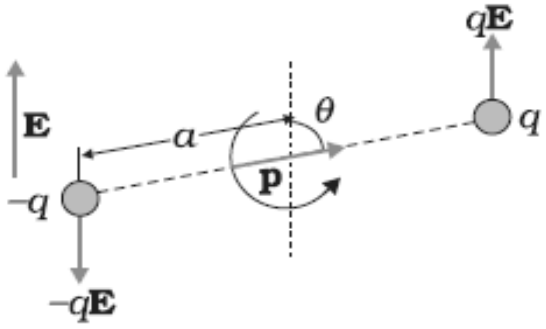


Fig. 4.15 Dipole in a uniform electric field.

There is a force $q\mathbf{E}$ on q and a force $-q\mathbf{E}$ on $-q$. The net force on the dipole is zero, since **E** is uniform. However, the charges are separated, so the forces act at different points, resulting in a torque on the dipole. When the net force is zero, the torque (couple) is independent of the origin. Its magnitude equals the magnitude of each force multiplied by the arm of the couple (perpendicular distance between the two antiparallel forces).

Magnitude of torque = $qE \times 2a \sin\theta$

$$= 2qa\, E \sin\theta$$

Its direction is normal to the plane of the paper, coming out of it.

The magnitude of $\boldsymbol{p \times E}$ is also $pE \sin\theta$ and its direction is normal to the paper, coming out of it. Thus,

$$\boldsymbol{\tau = p \times E} \qquad \textbf{(}4.22\textbf{)}$$

This torque will tend to align the dipole with the field **E**. When **p** is aligned with **E**, the torque is zero.

What happens if the field is not uniform? In that case the net force will evidently be non-zero. In addition there will, in general, be a torque on the system as before. The general case is involved, so let us consider the simpler situations when p is parallel to **E** or antiparallel to **E**. In either case, the net torque is zero, but there is a net force on the dipole if **E** is not uniform.

Figure 4.16 is self-explanatory. It is easily seen that when **p** is parallel to **E**, the dipole has a net force in the direction of increasing field. When **p** is

antiparallel to **E**, the net force on the dipole is in the direction of decreasing field. In general, the force depends on the orientation of **p** with respect to **E**.

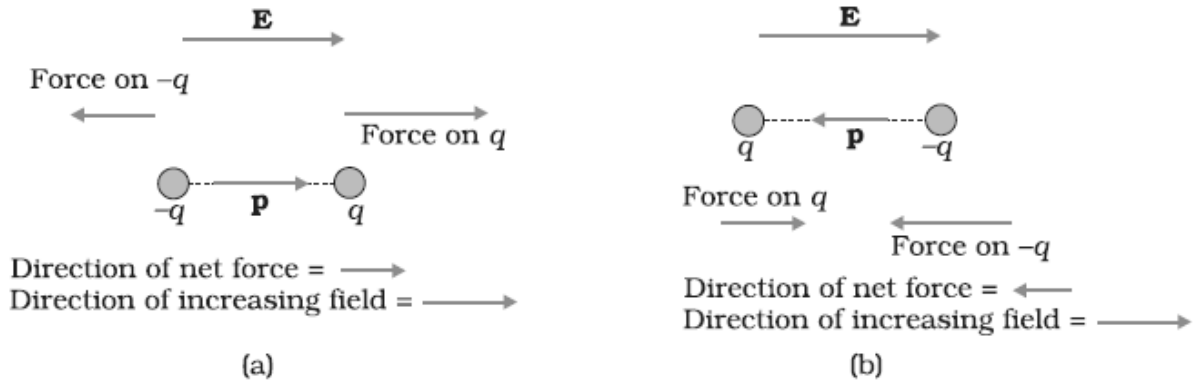


Fig.4.16 Electric force on a dipole: (a) **E** parallel to **p**, (b) **E** antiparallel to **p**.

This brings us to a common observation in frictional electricity. A comb run through dry hair attracts pieces of paper. The comb, as we know, acquires charge through friction. But the paper is not charged. What then explains the attractive force? Taking the clue from the preceding discussion, the charged comb 'polarizes' the piece of paper, i.e., induces a net dipole moment in the direction of field. Further, the electric field due to the comb is not uniform. In this situation, it is easily seen that the paper should move in the direction of the comb!

4.13 CONTINUOUS CHARGE DISTRIBUTION

We have so far dealt with charge configurations involving discrete charges $q_1, q_2, ..., q_n$. One reason why we restricted to discrete charges is that the mathematical treatment is simpler and does not involve calculus. For many purposes, however, it is impractical to work in terms of discrete charges and we need to work with continuous charge distributions. For example, on the surface of a charged conductor, it is impractical to specify the charge distribution in terms of the locations of the microscopic charged constituents. It is more feasible to consider an area element ΔS (Fig. 4.17) on the surface of the conductor (which is very small on the macroscopic scale but big enough to include a very large number of electrons) and specify the charge ΔQ on that element.

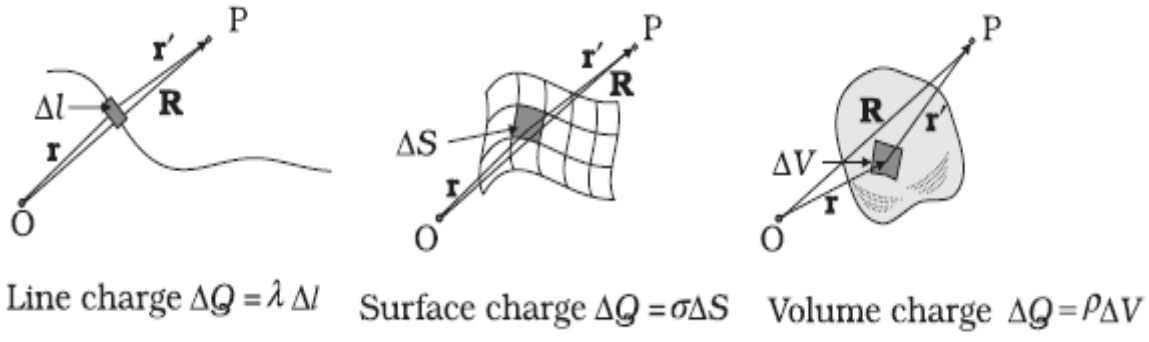


Fig. 4.17 Definition of linear, surface and volume charge densities. In each case, the $(\Delta l, \Delta S, \Delta V)$ chosen is small on the macroscopic scale but contains a very large number of microscopic constituents.

We then define a surface charge density σ at the area element by

$$\sigma = \frac{\Delta Q}{\Delta S} \quad (4.23)$$

We can do this at different points on the conductor and thus arrive at a continuous function σ, called the surface charge density. The surface charge density σ so defined ignores the quantisation of charge and the discontinuity in charge distribution at the microscopic level*. σ represents macroscopic surface charge density, which in a sense, is a smoothed out average of the microscopic charge density over an area element ΔS which as said before, is large microscopically but small macroscopically. The units for σ are C/m^2.

Similar considerations apply for a line charge distribution and a volume charge distribution. The *linear charge density* λ of a wire is defined by

$$\lambda = \frac{\Delta Q}{\Delta l} \quad (4.24)$$

where Δl is a small line element of wire on the macroscopic scale that, however, includes a large number of microscopic charged constituents, and ΔQis the charge contained in that line element. The units for λ are C/m. The volume charge density (sometimes simply called charge density) is defined in a similar manner:

$$\rho = \frac{\Delta Q}{\Delta V} \quad (4.25)$$

Where ΔQ is the charge included in the macroscopically small volume element ΔV that includes a large number of microscopic charged constituents. The units for ρ are C/m^3.

The notion of continuous charge distribution is similar to that we adopt for continuous mass distribution in mechanics. When we refer to the density of a liquid, we are referring to its macroscopic density. We regard it as a continuous fluid and ignore its discrete molecular constitution.

The field due to a continuous charge distribution can be obtained in much the same way as for a system of discrete charges, Eq. (4.10). Suppose a continuous charge distribution in space has a charge density ρ. Choose any convenient origin O and let the position vector of any point in the charge distribution be $\boldsymbol{r}$. The charge density ρ may vary from point to point, i.e., it is a function of $\boldsymbol{r}$. Divide the charge distribution into small volume elements of size ΔV. The charge in a volume element ΔV is ρΔV.

Now, consider any general point P (inside or outside the distribution) with position vector **R** (Fig. 4.17). Electric field due to the charge ρΔV is given by Coulomb's law:

$$\Delta E = \frac{1}{4\pi\varepsilon_0}\frac{\rho\Delta V}{r'^2}\,\hat{r'} \quad (4.26)$$

where r' is the distance between the charge element and P, and $\hat{r'}$ is a unit vector in the direction from the charge element to P. By the superposition principle, the total electric field due to the charge distribution is obtained by summing over electric fields due to different volume elements:

---

* At the At the microscopic level, charge distribution is discontinuous, because they are discrete charges separated by intervening space where there is no charge.

$$\mathbf{E} \cong \frac{1}{4\pi\varepsilon_0} \sum_{all\ \Delta v} \frac{\rho \Delta V}{r'^2} \hat{r'} \qquad (4.27)$$

Note that ρ, r', $\hat{r'}$all can vary from point to point. In a strict mathematical method, we should let $\Delta V \to 0$ and the sum then becomes an integral; but we omit that discussion here, for simplicity. In short, using Coulomb's law and the superposition principle, electric field can be determined for any charge distribution, discrete or continuous or part discrete and part continuous.

4.14 GAUSS'S LAW

As a simple application of the notion of electric flux, let us consider the total flux through a sphere of radius r, which encloses a point charge q at its centre. Divide the sphere into small area elements, as shown in Fig. 4.18.

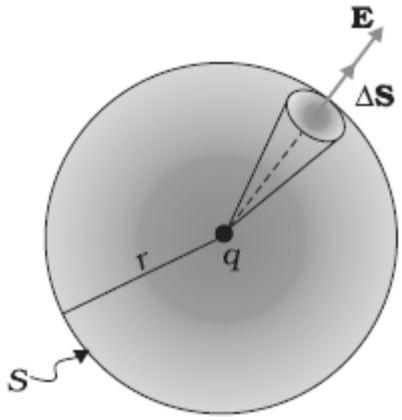


Fig. 4.18 Flux through a sphere enclosing a point charge q at its centre.

The flux through an area element $\Delta \boldsymbol{S}$ is

$$\Delta \emptyset = \boldsymbol{E}.\Delta \boldsymbol{S} = \frac{q}{4\pi\varepsilon_0 r^2} \hat{\boldsymbol{r}}.\Delta \boldsymbol{S} \qquad (4.28)$$

where we have used Coulomb's law for the electric field due to a single charge q. The unit vector $\hat{\boldsymbol{r}}$ is along the radius vector from the centre to the area element. Now, since the normal to a sphere at every point is along the radius vector at that point, the area element $\Delta \mathbf{S}$ and $\hat{\boldsymbol{r}}$ the same direction. Therefore,

$$\Delta \emptyset = \frac{q}{4\pi\varepsilon_0 r^2} \Delta \qquad (4.29)$$

since the magnitude of a unit vector is 1.

The total flux through the sphere is obtained by adding up flux through all the different area elements:

$$\emptyset = \sum_{all\ \Delta S} \frac{q}{4\pi\varepsilon_0 r^2} \Delta S$$

Since each area element of the sphere is at the same distance r from the charge,

$$\emptyset = \frac{q}{4\pi\varepsilon_0 r^2} \sum_{all\ \Delta S} \Delta S = \frac{q}{4\pi\varepsilon_0 r^2} S$$

Now S, the total area of the sphere, equals πr^2. Thus,

$$\emptyset = \frac{q}{4\pi\varepsilon_0 r^2}\, x\, 4\pi r^2 = \frac{q}{\varepsilon_0} \qquad (4.30)$$

Equation (4.30) is a simple illustration of a general result of electrostatics called Gauss's law.

We state *Gauss's law* without proof:

Electric flux through a closed surface S (4.31)

$= \frac{q}{\varepsilon_0}$

q = total charge enclosed by S

The law implies that the total electric flux through a closed surface is zero if no charge is enclosed by the surface. We can see that explicitly in the simple situation of Fig. 4.19.

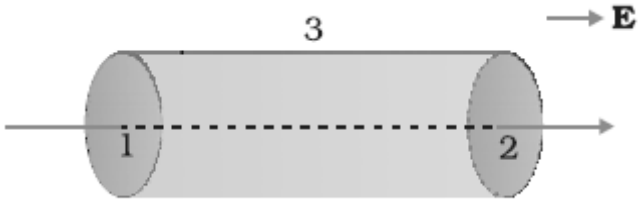


Fig. 4.19 Calculation of the flux of uniform electric field through the surface of a cylinder.

Here the electric field is uniform and we are considering a closed cylindrical surface, with its axis parallel to the uniform field **E.** The total flux $\emptyset$ through the surface is is $\emptyset = \emptyset_1 + \emptyset_2 + \emptyset_3$, where $\emptyset_1$and $\emptyset_2$represent the flux through the surfaces 1 and 2 (of circular cross-section) of the cylinder and $\emptyset_3$is the flux through the curved cylindrical part of the closed surface. Now the normal to the surface 3 at every point is perpendicular to **E**, so by definition of flux, $\emptyset_3 = 0$. Further, the outward normal to 2 is along **E** while the outward normal to 1 is opposite to E. Therefore,

$$\emptyset_1 = -ES_1, \qquad \emptyset_2 = +ES_2$$

$$S_1 = S_2 = S$$

where S is the area of circular cross-section. Thus, the total flux is zero, as expected by Gauss's law. Thus, whenever you find that the net electric flux through a closed surface is zero, we conclude that the total charge contained in the closed surface is zero.

The great significance of Gauss's law Eq. (4.31), is that it is true in general, and not only for the simple cases we have considered above. Let us note some important points regarding this law:

(i) Gauss's law is true for any closed surface, no matter what its shape or size

(ii) The term q on the right side of Gauss's law, Eq. (4.31), includes the sum of all charges enclosed by the surface. The charges may be located anywhere inside the surface.

(iii) In the situation when the surface is so chosen that there are some charges inside and some outside, the electric field [whose flux appears on the left side of Eq. (4.31)] is due to all the charges, both inside and outside S. The term q on the right side of Gauss's law, however, represents only the total charge inside S.

(iv) The surface that we choose for the application of Gauss's law is called the Gaussian surface. You may choose any Gaussian surface and apply Gauss's law. However, take care not to let the Gaussian surface pass through any discrete charge. This is because electric field due to a system of discrete charges is not well defined at the location of any charge. (As you go close to the charge, the field grows without any bound.) However, the Gaussian surface can pass through a continuous charge distribution.

(v) Gauss's law is often useful towards a much easier calculation of the electrostatic field *when the system has some symmetry*. This is facilitated by the choice of a suitable Gaussian surface.

(vi) Finally, Gauss's law is based on the inverse square dependence on distance contained in the Coulomb's law. Any violation of Gauss's law will indicate departure from the inverse square law.

4.15 APPLICATIONS OF GAUSS'S LAW

The electric field due to a general charge distribution is, as seen above, given by Eq. (4.27). In practice, except for some special cases, the summation (or integration) involved in this equation cannot be carried out to give electric field at every point in space. For some symmetric charge configurations, however, it is possible to obtain the electric field in a simple way using the Gauss's law. This is best understood by some examples.

4.15.1 Field due to an infinitely long straight uniformly charged wire

Consider an infinitely long thin straight wire with uniform linear charge density λ. The wire is obviously an axis of symmetry. Suppose we take the radial vector from O to P and rotate it around the wire. The points P, P', P'' so obtained are completely equivalent with respect to the charged wire. This implies that the electric field must have the same magnitude at these points. The direction of electric field at every point must be radial (outward if $\lambda > 0$, inward if $\lambda < 0$). This is clear from Fig. 4.20.

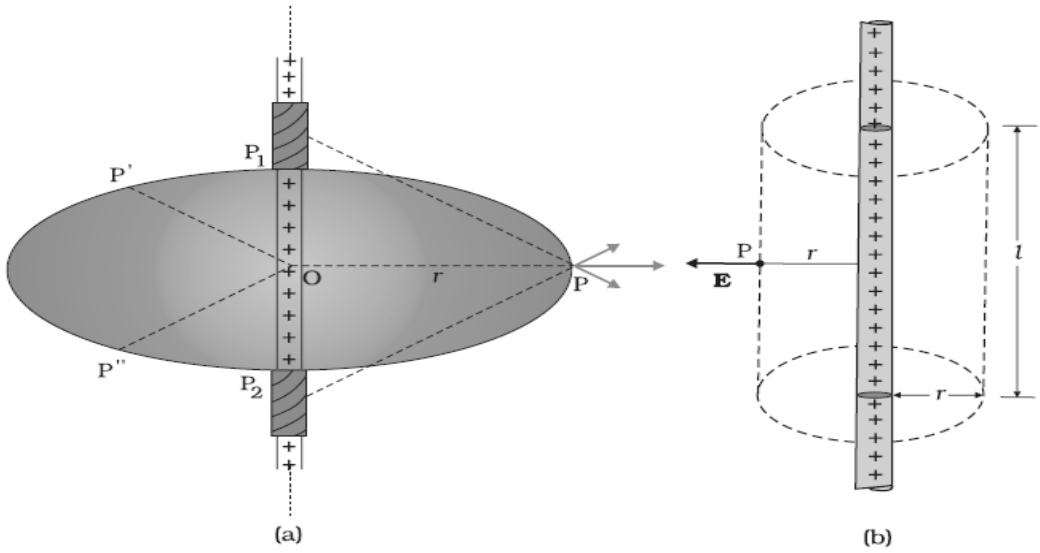


Fig. 4.20 (a) Electric field due to an infinitely long thin straight wire is radial, (b) The Gaussian surface for a long thin wire of uniform linear charge density.

Consider a pair of line elements P_1 and P_2 of the wire, as shown. The electric fields produced by the two elements of the pair when summed give a resultant electric field which is radial (the components normal to the radial vector cancel). This is true for any such pair and hence the total field at any point P is radial. Finally, since the wire is infinite, electric field does not depend on the position of P along the length of the wire. In short, the electric field is everywhere radial in the plane cutting the wire normally, and its magnitude depends only on the radial distance *r*.

To calculate the field, imagine a cylindrical Gaussian surface, as shown in the Fig. 4.20(b). Gaussian surface, as shown in the Fig. 4.20(b). Since the field is everywhere radial, flux through the two ends of the cylindrical Gaussian surface is zero. At the cylindrical part of the surface, **E** is normal to the surface at every point, and its magnitude is constant, since it depends only on *r*. The surface area of the curved part is $2\pi rl$, where l is the length of the cylinder.

Flux through the Gaussian surface

= flux through the curved cylindrical part of the surface

$= E \times 2\pi rl$

The surface includes charge equal to $\lambda\, l$.. Gauss's law then gives

$$E \times 2\pi rl = \frac{\lambda l}{\varepsilon_0}$$

i.e., $$E = \frac{\lambda}{2\pi\varepsilon_0 r} \qquad (4.32)$$

Vectorially, **E** at any point is given by

$$E = \frac{\lambda}{2\pi\varepsilon_0 r}\,\hat{n}$$

where $\hat{n}$ is the radial unit vector in the plane normal to the wire passing through the point. **E** is directed outward if λ is positive and inward if λ is negative.

Note that when we write a vector **A** as a scalar multipled by a unit vector, i.e., as **A** = A$\hat{a}$, the scalar A is an algebraic number. It can be negative or positive. The direction of A will be the same as that of the unit vector $\hat{a}$ if A>0 and opposite to $\hat{a}$ if A<0. When we want to restrict to non-negative values, we use the symbol $|A|$ and call it the modulus of **A.** Thus, $|A| \geq 0$.

Also note that though only the charge enclosed by the surface (λl) was included above, the electric field **E** is due to the charge on the entire wire. Further, the assumption that the wire is infinitely long is crucial. Without this assumption, we cannot take **E** to be normal to the curved part of the cylindrical Gaussian surface. However, Eq. (4.32) is approximately true for electric field around the central portions of a long wire, where the end effects may be ignored.

4.15.2 Field due to a uniformly charged infinite plane sheet

Let σ be the uniform surface charge density of an infinite plane sheet (Fig. 4.21). We take the x-axis normal to the given plane. By symmetry, the electric field will not depend on y and z coordinates and its direction at every point must be parallel to the x-direction.

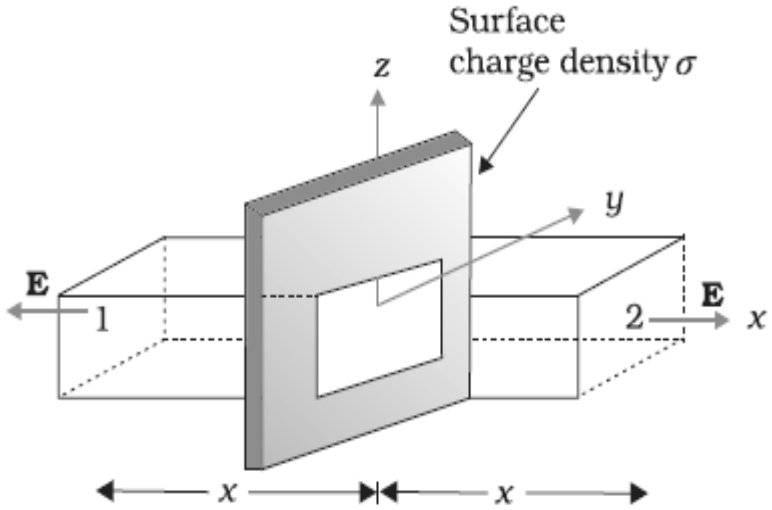


Fig. 4.21 Gaussian surface for a uniformly charged infinite plane sheet.

We can take the Gaussian surface to be a rectangular parallelepiped of cross sectional area A, as shown. (A cylindrical surface will also do.) As seen from the figure, only the two faces 1 and 2 will contribute to the flux; electric field lines are parallel to the other faces and they, therefore, do not contribute to the total flux.

The unit vector normal to surface 1 is in –x direction while the unit vector normal to surface 2 is in the +x direction. Therefore, flux $\mathbf{E}.\Delta\mathbf{S}$ through both the surfaces are equal and add up. Therefore the net flux through the Gaussian surface is 2 EA. The charge enclosed by the closed surface is σA. Therefore by Gauss's law,

$$2\,EA = \frac{\sigma A}{\varepsilon_0}$$

$$\text{Or, } E = \frac{\sigma}{2\varepsilon_0}$$

Vertically,

$$E = 2\frac{\sigma}{\varepsilon_0}\,\hat{n} \qquad (4.33)$$

where $\hat{n}$ is a unit vector normal to the plane and going away from it.

E is directed away from the plate if σ is positive and toward the plate if σ is negative. Note that the above application of the Gauss' law has brought out an additional fact: E is independent of x also.

For a finite large planar sheet, Eq. (4.33) is approximately true in the middle regions of the planar sheet, away from the ends.

4.15.3 Field due to a uniformly charged thin spherical shell

Let σ be the uniform surface charge density of a thin spherical shell of radius R (Fig. 4.22). The situation has obvious spherical symmetry. The field at any point P, outside or inside, can depend only on r (the radial distance from the centre of the shell to the point) and must be radial (i.e., along the radius vector).

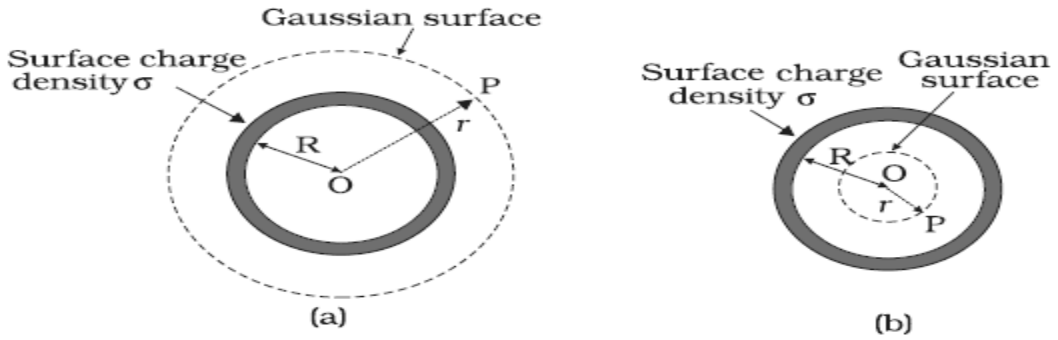


Fig.1.31 Gaussian surfaces for a point with (a) r > R, (b) r < R.

(i) ***Field outside the shell:*** Consider a point P outside the shell with radius vector r. To calculate **E** at P, we take the Gaussian surface to be a sphere of radius r and with centre O, passing through P. All points on this sphere are equivalent relative to the given charged configuration. (That is what we mean by spherical symmetry.) The electric field at each point of the Gaussian surface, therefore, has the same magnitude **E** and is along the radius vector at each point. Thus, E and ΔS at every point are parallel and the flux through each element is E ΔS. Summing over all ΔS, the flux through the Gaussian surface is E × 4 π r^2. The charge enclosed is σ × 4 πR^2. By Gauss's law

$$E\ x\ 4\pi r^2 = \frac{\sigma}{\varepsilon_0}\ 4\pi R^2$$

$$Or, E = \frac{\sigma R^2}{\varepsilon_0 r^2} = \frac{q}{4\pi\varepsilon_0 r^2}$$

where $q = 4\pi R^2\sigma$ is the total charge on the spherical shell.

Vertically,

$$E = \frac{q}{4\pi\varepsilon_0 r^2}\hat{r} \qquad (4.34)$$

The electric field is directed outward if q > 0 and inward if q < 0. This, however, is exactly the field produced by a charge q placed at the centre O. Thus for points outside the shell, the field due to a uniformly charged shell is as if the entire charge of the shell is concentrated at its centre.

(ii) Field inside the shell: In Fig. 4.22(b), the point P is inside the shell. The Gaussian surface is again a sphere through P centred at O.

The flux through the Gaussian surface, calculated as before, is E × 4πr^2. However, in this case, the Gaussian surface encloses no charge. Gauss's law then gives

$$E\ x\ 4\pi r^2 = 0$$

that is, the field due to a uniformly charged thin shell is zero at all points inside the shell*****. This important result is a direct consequence of Gauss's law which follows from Coulomb's law. The experimental verification of this result confirms the 1/r^2 dependence in Coulomb's law.

---

* Compare this with a uniform mass shell discussed in Section 8.5 of Class XI Textbook of Physics.

ON SYMMETRY OPERATIONS

In Physics, we often encounter systems with various symmetries. Consideration of these symmetries helps one arrive at results much faster than otherwise by a straightforward calculation. Consider, for example an infinite uniform sheet of charge (surface charge density σ) along the y-z plane. This system is unchanged if (a) translated parallel to the y-z plane in any direction, (b) rotated about the x-axis through any angle. As the system is unchanged under such symmetry operation, so must its properties be. In particular, in this example, the electric field **E** must be unchanged.

Translation symmetry along the y-axis shows that the electric field must be the same at a point (0, y_1, 0) as at (0, y_2, 0). Similarly translational symmetry along the z-axis shows that the electric field at two point (0, 0, z_1) and (0, 0, z_2) must be the same. By using rotation symmetry around the x-axis, we can conclude that E must be perpendicular to the y-z plane, that is, it must be parallel to the x-direction.

Try to think of a symmetry now which will tell you that the magnitude of the electric field is a constant, independent of the x-coordinate. It thus turns out that the magnitude of the electric field due to a uniformly charged infinite conducting sheet is the same at all points in space. The direction, however, is opposite of each other on either side of the sheet.

Compare this with the effort needed to arrive at this result by a direct calculation using Coulomb's law.

SUMMARY

1. Electric and magnetic forces determine the properties of atoms, molecules and bulk matter.
2. From simple experiments on frictional electricity, one can infer that there are two types of charges in nature; and that like charges repel and unlike charges attract. By convention, the charge on a glass rod rubbed with silk is positive; that on a plastic rod rubbed with fur is then negative.
3. Conductors allow movement of electric charge through them, insulators do not. In metals, the mobile charges are electrons; in electrolytes both positive and negative ions are mobile.
4. Electric charge has three basic properties: quantisation, additivity and conservation.

   Quantisation of electric charge means that total charge (q) of a body is always an integral multiple of a basic quantum of charge (e) i.e., q = n e, where n = 0, ±1, ±2, ±3, ... Proton and electron have charges +e, –e, respectively. For macroscopic charges for which n is a very large number, quantisation of charge can be ignored.

   Additivity of electric charges means that the total charge of a system is the algebraic sum (i.e., the sum taking into account proper signs) of all individual charges in the system.

   Conservation of electric charges means that the total charge of an isolated system remains unchanged with time. This means that when bodies are charged through friction, there is a transfer of electric charge from one body to another, but no creation or destruction of charge.
5. Coulomb's Law: The mutual electrostatic force between two point charges

q_1 and q_2 is proportional to the product q_1q_2 and inversely proportional to the square of the distance r_{21} separating them. Mathematically,

F_{21} = force on q_2 due to $q1 = \frac{k\,(q_1 q_2)}{r_{21}^2}\,\hat{r}_{21}$

where $\hat{r}_{21}$is a unit vector in the direction from q_1 to q_2 and $k = \frac{1}{4\pi\varepsilon_0}$ is the constant of proportionality.

In SI units, the unit of charge is coulomb. The experimental value of the constant ε_0 is

$$\varepsilon_0 = 8.854\ x\ 10 - 12\ c^2\ N^{-1}m^{-2}$$

The approximate value of k is

$$k = 9\ x\ 10^9 N\ m^2 C^{-2}$$

6. The ratio of electric force and gravitational force between a proton and an electron is

$$\frac{k\ e^2}{G\ m_e m_p} \cong 2.4\ x\ 10^{39}$$

7. *Superposition Principle:* The principle is based on the property that the forces with which two charges attract or repel each other are not affected by the presence of a third (or more) additional charge(s). For an assembly of charges q_1, q_2, q_3, ..., the force on any charge, say q_1, is the vector sum of the force on q_1 due to q_2, the force on q_1 due to q_3, and so on. For each pair, the force is given by the Coulomb's law for two charges stated earlier.
8. The electric field **E** at a point due to a charge configuration is the force on a small positive test charge q placed at the point divided by the magnitude of the charge. Electric field due to a point charge q has a magnitude $\frac{|q|}{4\pi\varepsilon_0 r^2}$; it is radially outwards from q, if q is positive, and radially inwards if q is negative. Like Coulomb force, electric field also satisfies superposition principle.
9. An electric field line is a curve drawn in such a way that the tangent at each point on the curve gives the direction of electric field at that point. The relative closeness of field lines indicates the relative strength of electric field at different points; they crowd near each other in regions of strong electric field and are far apart where the electric field is weak. In regions of constant electric field, the field lines are uniformly spaced parallel straight lines.
10. Some of the important properties of field lines are: (i) Field lines are continuous curves without any breaks. (ii) Two field lines cannot cross each other. (iii) Electrostatic field lines start at positive charges and end at negative charges —they cannot form closed loops.
11. An electric dipole is a pair of equal and opposite charges q and –q separated by some distance $2a$. Its dipole moment vector **p** has magnitude $2qa$ and is in the direction of the dipole axis from $-q$ to q.
12. Field of an electric dipole in its equatorial plane (i.e., the plane perpendicular to its axis and passing through its centre) at a distance r from the centre:

$$E = \frac{-\boldsymbol{p}}{4\pi\varepsilon_0} \frac{1}{(a^2+r^2)^{\frac{3}{2}}}$$

$$= \frac{-p}{4\pi\varepsilon_0 r^3}, \qquad \text{for} \quad r >> a$$

Dipole electric field on the axis at a distance r from the centre:

$$E = \frac{2pr}{4\pi\varepsilon_0(r^2 - a^2)^2}$$

$$\cong \frac{2p}{4\pi\varepsilon_0 r^3} \qquad for\ r >> a$$

The $1/r^3$ dependence of dipole electric fields should be noted in contrast to the $1/r^2$ dependence of electric field due to a point charge.

13. In a uniform electric field E, a dipole experiences a torque τ given by

$$\boldsymbol{\tau = p\ x\ E}$$

but experiences no net force.

14. The flux $\Delta\emptyset$ of electric field E through a small area element ΔS is given by

$$\Delta\emptyset = \mathbf{E}.\Delta\mathbf{S}$$

The vector area *element* $\Delta\boldsymbol{S}$ is

$$\Delta\boldsymbol{S} = \Delta S\,\hat{n}$$

where ΔS is the magnitude of the area element and $\hat{n}$ is normal to the area element, which can be considered planar for sufficiently small ΔS.

For an area element of a closed surface, $\hat{\boldsymbol{n}}$ is taken to be the direction of outward normal, by convention.

15. *Gauss's law:* The flux of electric field through any closed surface S is $1/\varepsilon_0$ times the total charge enclosed by S. The law is especially useful in determining electric field **E**, when the source distribution has simple symmetry:

(i) Thin infinitely long straight wire of uniform linear charge density λ

$$E = \frac{\lambda}{2\pi\varepsilon_0 r}\,\hat{n}$$

where r is the perpendicular distance of the point from the wire and $\hat{n}$ is the radial unit vector in the plane normal to the wire passing through the point.

(ii) *Infinite thin plane sheet of uniform surface charge density* σ

$$E = \frac{\sigma}{2\varepsilon_0}\hat{n}$$

where $\hat{n}$ is a unit vector normal to the plane, outward on either side.

(iii) Thin spherical shell of uniform surface charge density σ

$$E = \frac{q}{4\pi\varepsilon_0 r^2}\,\hat{r} \qquad (r \geq R)$$

$$E = 0 \qquad (r \geq R)$$

where r is the distance of the point from the centre of the shell and R the radius of the shell. q is the total charge of the shell: $q = 4\pi R^2\sigma$. The electric field outside the shell is as though the total charge is concentrated at the centre. The same result is true for a solid sphere of uniform volume charge density. The field is zero at all points inside the shell.

Physical quantity	Symbol	Dimensions	Unit	Remarks
Vector area element	$\Delta \mathbf{S}$	$[L^2]$	m^2	$\Delta \mathbf{S} = \Delta S\, \hat{\mathbf{n}}$
Electric field	$\mathbf{E}$	$[MLT^{-3}A^{-1}]$	$V\ m^{-1}$	
Electric flux	ϕ	$[ML^3\ T^{-3}A^{-1}]$	V m	$\Delta\phi = \mathbf{E}\cdot\Delta\mathbf{S}$
Dipole moment	$\mathbf{p}$	[LTA]	C m	Vector directed from negative to positive charge
Charge density				
linear	λ	$[L^{-1}\ TA]$	$C\ m^{-1}$	Charge/length
surface	σ	$[L^{-2}\ TA]$	$C\ m^{-2}$	Charge/area
volume	ρ	$[L^{-3}\ TA]$	$C\ m^{-3}$	Charge/volume

VERY SHORT ANSWER QUESTIONS (2 MARKS)

1. What is meant by the statement 'charge is quantized'?
2. Repulsion is the sure test of charging than attraction. Why?
3. How many electrons constitute 1 C of charge?
4. What happens to the weight of a body when it is charged positively ?
5. What happens to the force between two charges if the distance between them is
   (a) halved (b) doubled ?
6. The electric lines of force do not intersect .Why ?
7. State Gauss's law in electrostatics.
8. When is the electric flux negative and when is it positive ?
9. Write the expression for electric intensity due to an infinite long charged wire at a distance radial distance r from the wire.
10. Write the expression for electric intensity due to an infinite plane sheet of charge.
11. Write the expression for electric intensity due to a charged conducting spherical shell at points outside and inside the shell.

SHORT ANSWER QUESTIONS (4 MARKS)

1. State and explain Coulomb's inverse square law in electrically.
2. Define intensity of electric field at a point. Derive an expression for the intensity due to a point charge.
3. Derive the equation for the couple acting on a electric dipole in a uniform electric field.

4. Derive an expression for the intensity of the electric field at a point on the axial line of an electric dipole.
5. Derive an expression for the intensity of the electric field at a point on the equatorial plane of an electric dipole.
6. State Gauss's law in electrostatics and explain its importance.

LONG ANSWER QUESTIONS (8 MARKS)

1. Define electric flux. Applying Gauss's law and derive the expression for electric intensity due to an infinite long straight charged wire. (Assume that the electric field is everywhere radial and depends only on the radial distance r of the point from the wire.)
2. State Gauss's law in electrostatics. Applying Gauss's law derive the expression for electric intensity due to an infinite plane sheet of charge.
3. Applying Gauss's law derive the expression for electric intensity due to a charged conducting spherical shell at (i) a point outside the shell (ii) a point on the surface of the shell and (iii) a point inside the shell.

Chapter 5

ELECTROSTATIC POTENTIAL AND CAPACITANCE

5. 1 INTRODUCTION

In Chapter 6 and 8 (Class XI), the notion of potential energy was introduced. When an external force does work in taking a body from a point to another against a force like spring force or gravitational force, that work gets stored as potential energy of the body. When the external force is removed, the body moves, gaining kinetic energy and losing an equal amount of potential energy. The sum of kinetic and potential energies is thus conserved. Forces of this kind are called conservative forces. Spring force and gravitational force are examples of conservative forces.

Coulomb force between two (stationary) charges is also a conservative force. This is not surprising, since both have inverse-square dependence on distance and differ mainly in the proportionally constants – the masses in the gravitational law are replaced by charges in Coulomb's law. Thus, like the potential energy of a mass in a gravitational field. We can define electrostatic potential energy of a charge in an electrostatic field.

Consider an electrostatic field E due to some charge configuration. First, for simplicity, consider the field E due to a charge Q placed at the origin. Now, imagine that we bring a test charge q from a point R to a point P against the repulsive force on it due to the charge Q. With reference to Fig. 5.1 , this will happen if Q and q are both positive or both negative. For definiteness, let us take Q , q > 0.

Fig. 2.1 A test charge *q* (> 0) is moved from the point R to the point P against the repulsive force on it by the charge *Q* (> 0) placed at the origin

Two remarks may be made here. First, we assume that the original configuration, namely the charge Q at the origin (or else, we keep Q fixed at the origin by some unspecified force). Second , in bringing the charge q from R to P, we apply an external force $\mathbf{F}_{ext}$ just enough to counter the repulsive electric force $\mathbf{F}_E$ **(i.e,** $\mathbf{F}_{ext} = -\mathbf{F}_E$**)**. This means there is no net force on or acceleration of the charge q when it is brought from R to P, i.e., it is brought with infinitesimally slow constant speed. In this situation, work done by the external force is the negative of the work done by the electric force, and gets fully stored in the form of potential energy of the charge q. If the external force is removed on reaching P, the electric force will take the charge away from Q – the stored energy (potential energy) at P is used to provide

kinetic energy to the charge q in such a way that the sum of the kinetic and potential energies is conserved.

Thus, work done by external forces in moving a charge q from R to P is

$$W_{RP} = \int_R^P F_{ext} \cdot dr$$

$$= -\int_R^P F_E \cdot dr \qquad (5.1)$$

This work done is against electrostatic repulsive force and gets stored as potential energy

At every point in electric field, a particle with charge q possesses a certain electrostatic potential energy, this work done increases its potential energy by an amount equal to potential energy difference between points R and P.

Thus , potential energy difference

$$\Delta U = U_P - U_R = W_{RP} \qquad (5.2)$$

(Note here that this displacement is in an opposite sense to the electric force and hence work done by electric field is negative, i.e., -W_{RP}.)

Therefore, we can define electric potential energy difference between two points as the work required to be done by an external force in moving (without accelerating) charge q from one point to another for electric field of any arbitrary charge configuration.

Two important comments may be made at this stage:

(i) The right side of Eq.(5.2) depends only on the initial and final positions of the charge. It means that the work done by an electrostatic field in moving a charge from one point to another depends only on the initial and the final points and is independent of the path taken to go from one point to the other. This is the fundamental characteristic of a conservative force. The concept of the potential energy would not be meaningful if the work depends on the path. The path-independence of work done by an electrostatic field can be proved using the Coulomb's law. We omit this proof here.

(ii) Equation (5.2) defines potential energy difference in terms of the physically meaningful quantity work. Clearly, potential energy so defined is undetermined to within an additive constant. What this means is that the actual value of potential energy is not physically significant: it is only the difference of potential energy that is significant.

We can always add an arbitrary constant α to potential energy at every point. Since this will not change the potential energy difference:

$$(U_P + \alpha) - (U_R + \alpha) = U_P - U_R$$

Put it differently, there is a freedom in choosing the point where potential energy is zero. A choice is to have electrostatic potential energy zero at infinity. With this choice, if we take the point R at infinity, we get from Eq.

$$W_{\infty P} = U_P - U_\infty = U_P \qquad (5.3)$$

Since the point P is arbitrary Eq. provides us with a definition of potential energy of a charge q at any point. Potential energy of charge q at a point (in the presence of field due

to any charge configuration) is the work done by the external force (equal and opposite to the electric force) in bringing the charge q from infinity to that point.

Count Alessandro Volta (1745 – 1827)

Italian physicist, professor at Pavia. Volta established that the *animal electricity* observed by Luigi Galvani, 1737–1798, in experiments with frog muscle tissue placed in contact with dissimilar metals, was not due to any exceptional property of animal tissues but was also generated whenever any wet body was sandwiched between dissimilar metals. This led him to develop the first *voltaic pile*, or battery, consisting of a large stack of moist disks of cardboard (electrolyte) sandwiched between disks of metal (electrodes).

5.2 ELECTROSTATIC POTENTIAL

Consider any general static charge configuration. We define potential energy of a test charge q in terms of the work done on the charge q. This work is obviously proportional to q. Since the force at any point is qE. Where E is the electric field at that point due to the given charge configuration. It is, therefore, convenient to divide the work by the amount of charge q. So that the resulting quantity is independent of q. In other words, work done per unit test charge is characteristic of the electric field associated with the charge configuration. This leads to the idea of electrostatic potential V due to a given charge configuration. From Eq.(5.1) , we get :

Work done by external force in bringing a unit positive charge from point R to P

$$= V_P - V_R \left(= \frac{U_P - U_R}{q} \right) \qquad (5.4)$$

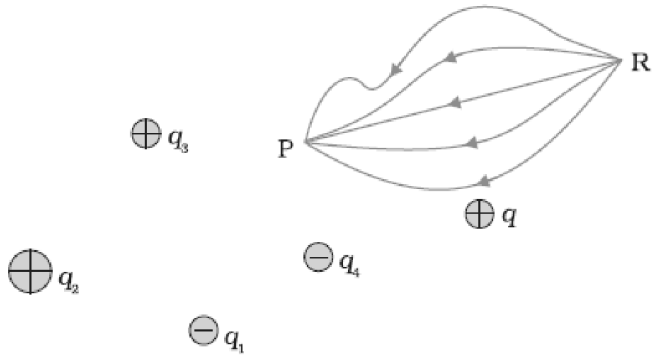


Fig. 5.2 Work done on a test charge *q* by the electrostatic field due to any given charge configuration is independent of the path, and depends only on its initial and final positions.

Where V_P and V_R are the electrostatic potentials at P and R, respectively. Note, as before , that it is not the actual value of potential but the potential difference that is physically significant. If, as before, we choose the potential to be zero at infinity. Eq. (5.4) Implies :

Work done by an external force in bringing a unit positive charge from infinity to a point = electrostatic potential (V) at that point.

In other words, the electrostatic potential (V) at any point in a region with electrostatic field is the work done in bringing a unit positive charge (without acceleration) from infinity to that point.

The qualifying remarks made earlier regarding potential energy also apply to the definition of potential. To obtain the work done per unit test charge, we should take an infinitesimal test charge δq. obtain the work done δW in bringing it from infinity to the point and determine the ratio $\delta W / \delta q$. Also, the external force at every point of the path is to be equal and opposite to the electrostatic force on the test charge at that point.

5.3 POTENTIAL DUE TO A POINT CHARGE

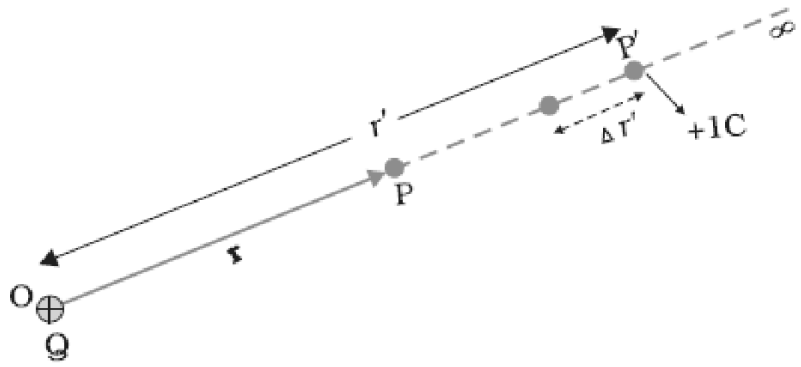


Fig. 5.3 Work done in bringing a unit positive test charge from infinity to the point P, against the repulsive force of charge Q (Q > 0), is the potential at P due to the charge Q.

Consider a point charge Q at the origin (Fig 5.3) . For definiteness, take Q to be positive. We wish to determine the potential at any point P with position vector **r** from the origin. For that we must calculate the work done in bringing a unit positive test charge from infinity to the point P. For Q > 0, the work done against the repulsive force on the test charge is positive. Since work done is independent of the path, we choose a convenient path-along the radial direction from infinity to the point P.

At some intermediate point P on the path , the electrostatic force on a unit positive charge is $\frac{Q \times 1}{4\pi\varepsilon_0 r'^2}\hat{r}'$ (5.5)

Where $\hat{r}'$ is the unit vector along OP' . Work done against this force from **r'** to **r' +Δr'** is

$$\Delta W = -\frac{Q}{4\pi\varepsilon_0 r'^2}\Delta r' \qquad (5.6)$$

The negative sign appears because for $\Delta r' < 0$, ΔW is positive. Total work done (W) by the external force is obtained by integrating Eq.(5.6) From $r' = \infty \; to \; r' = r$,

$$W = -\int_{\infty}^{r} \frac{Q}{4\pi\varepsilon_0 r'^2}\, dr' = \frac{Q}{4\pi\varepsilon_0 r'} \int_{\infty}^{r} = \frac{Q}{4\pi\varepsilon_0 r} \quad \textbf{(5.7)}$$

This , by definition is the potential at P due to the charge Q

$$V(r) = \frac{Q}{4\pi\varepsilon_0 r} \quad \textbf{(5.8)}$$

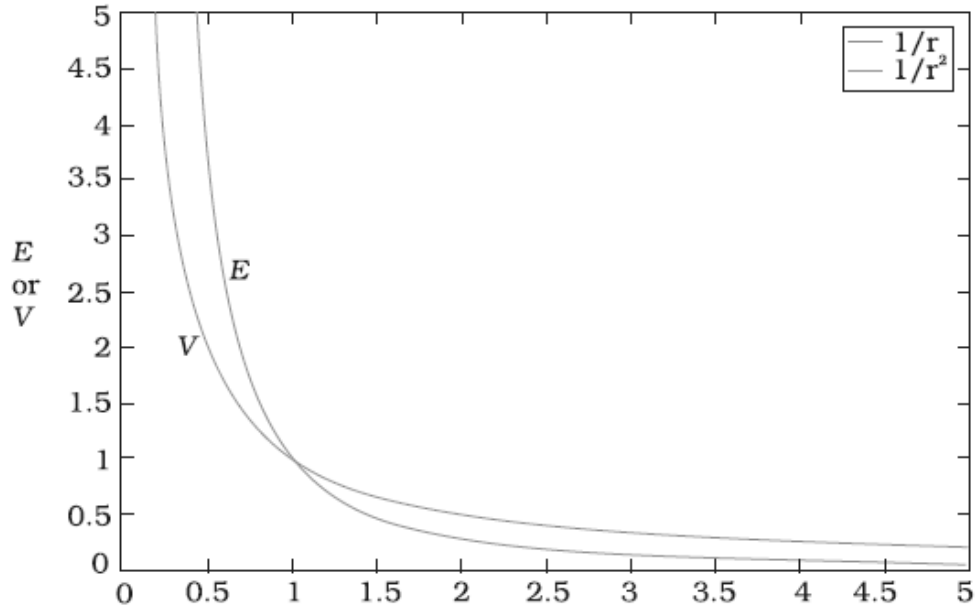


Fig. 5.4 Variation of potential *V* with *r* [in units of ($Q/4\pi\varepsilon 0$) m^{-1}] (blue curve) and field with *r* [in units of ($Q/4\pi\varepsilon 0$) m^{-2}] (black curve) for a point charge *Q*.

Equation is true for any sign of the charge Q, though we considered Q > 0 in its derivation. For Q < 0, V < 0 , i. e., work done (by the external force) per unit positive test charge in bringing it from infinity to the point is negative. This is equivalent to saying that work done by the electrostatic force in bringing the unit positive charge form infinity to the point P is positive.[This is an it should be, since for Q < 0, the force on a unit positive test charge is attractive , so that the electrostatic force and the displacement (from infinity to P) are in the same direction.] Finally, we note that Eq. is consistent with the choice that potential at infinity be zero.

Fig. 5.4 Shows how the electrostatic potential ($\propto 1/r$) and the electrostatic field ($\propto 1/r^2$) varies with r.

5.4 POTENTIAL DUE TO AN ELECTRIC DIPOLE

As we learnt in the last chapter, an electric dipole consists of two charges q and –q separated by a (small) distance 2a. Its total charge is zero. It is characterised by a dipole moment vector **P** whose magnitude is q x 2a and which points in the direction from –q to q . We also saw that the electric field of a dipole at a point with position vector **r** depends not just on the magnitude r, but also on the angle between **r** and **p**. Further, the field falls off, at large distance, not as $1/r^2$ (typical of field due to a single charge) but as $1/r^3$. We, now, determine the electric potential due to a dipole and contrast it with the potential due to a single charge.

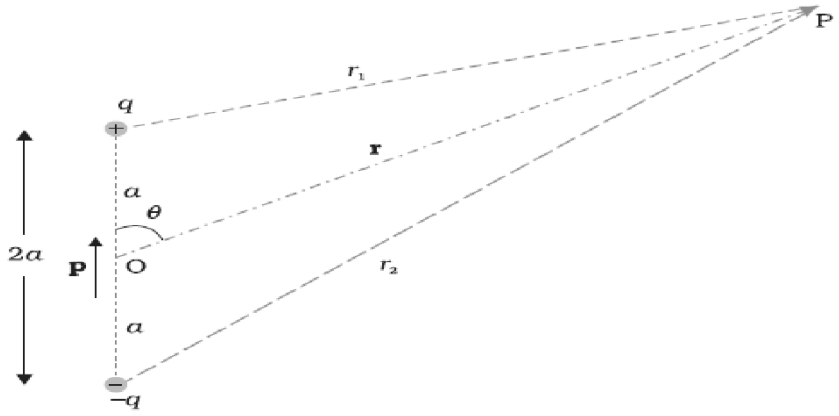


Fig. 5.5 Quantities involved in the calculation of potential due to a dipole.

As before, we take the origin at the centre of the dipole. Now we know that the electric field obey the superposition principle. Since [potential is related to the work done by the field, electrostatic potential also follows the superposition principle. Thus, the potential due to the dipole is the sum of potentials due to the charges q and –q

$$.V = \frac{1}{4\pi\varepsilon_0}\left(\frac{q}{r_1} - \frac{q}{r_2}\right) \quad (5.9)$$

Where r_1 and r_2 are the distances of the point P from q and –q, respectively.

Now, by geometry,

$$r_1^2 = r^2 + a^2 - 2ar\cos\theta$$

$$r_2^2 = r^2 + a^2 + 2ar\cos\theta \quad (5.10)$$

We take r much greater than a (r >> a) and retain terms only up to the first order in a/r

$$r_1^2 = r^2\left(1 - \frac{2a\cos\theta}{r} + \frac{a^2}{r^2}\right)$$

$$\cong r^2\left(1 - \frac{2a\cos\theta}{r}\right) \quad (5.11)$$

Similarly,

$$r_2^2 \cong r^2\left(1 + \frac{2a\cos\theta}{r}\right) \quad (5.12)$$

Using the Binomial theorem and retaining terms up to the first order in a / r:

we obtain ,

$$\frac{1}{r_1} \cong \frac{1}{r}\left(1 - \frac{2a\cos\theta}{r}\right)^{-1/2} \cong \frac{1}{r}\left(1 + \frac{a}{r}\cos\theta\right) \quad [5.13(a)]$$

$$\frac{1}{r_2} \cong \frac{1}{r}\left(1 + \frac{2a\cos\theta}{r}\right)^{-1/2} \cong \frac{1}{r}\left(1 - \frac{a}{r}\cos\theta\right) \quad [5.13(b)]$$

Using Eqs. (0.9) and (0.13) and p = 2qa , we get

$$V = \frac{q}{4\pi\varepsilon_0}\frac{2a\cos\theta}{r^2} = \frac{p\cos\theta}{4\pi\varepsilon_0 r^2} \quad (5.14)$$

Now, p cos θ = **p. r**

Where $\hat{r}$ is the unit vector along the position vector **OP**.

The electric potential of a dipole is then given by

$$V = \frac{1}{4\pi\varepsilon_0}\frac{p.\hat{r}}{r^2} : (\text{r} >> \text{a}) \qquad (5.15)$$

Eq (5.15) is , an indicated, approximately true only for distances large compared to the size of the dipole. So that higher order terms in a / r are negligible . For a point dipole **P** at the origin . Eq(5.15) is, however , exact.

From Eq.(5.15) , potential on the dipole axis ($\theta = 0.\pi$) is given by

$$V = \pm\frac{1}{4\pi\varepsilon_0}\frac{p}{r^2} \qquad (5.16)$$

[Positive sign for θ = 0, negative sign for $\theta = \pi$]. The potential in the equatorial plane ($\theta = \pi/2$) is zero.

The important contrasting features of electric potential of a dipole from that due to a single charge are clear from Eqs. (5.8) and (5.15):

(i) The potential due to a dipole depends not just on r but also on the angle between the position vector **r** and the dipole moment vector **p**.

(It is, however, axially symmetric about p. That is, if you rotate the position vector **r** about **p**, keeping θ fixed, the points corresponding to P on the cone so geberated will have the same potential as at P.)

(ii) The electric dipole potential falls off, at large distance, as $1/r^2$, not as 1/r, characteristic of the potential due to a single charge, (You can refer to the Fig. 5.5 for graphs of $1 / r^2$ versus r and 1 / r versus r, drawn there in another context.)

5.5 POTENTIAL DUE TO A SYSTEM OF CHARGES

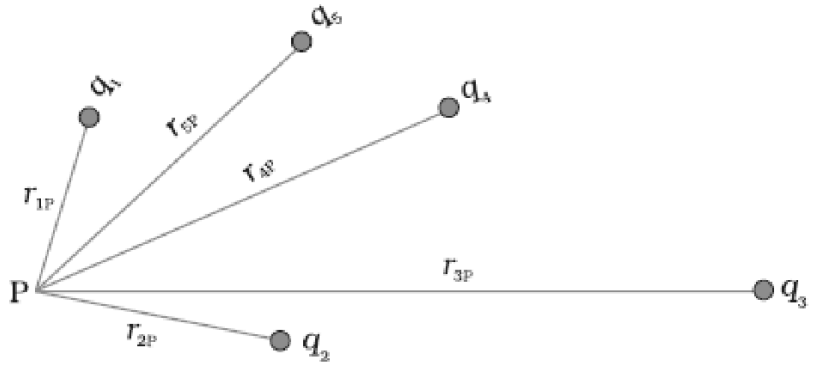


Fig. 5.6 Potential at a point due to a system of charges is the sum of potentials due to individual charges.

Consider a system of charges q_1, q_2 ..... q_n with position vectors r_1 , r_2.....r_n relative to some origin (Fig 5.6) . The potential V_1 at P due to the charge q_1 is

$$V_1 = \frac{1}{4\pi\varepsilon_0}\frac{q_1}{r_{1p}}$$

Where r_{1p} is the distance between q_1 and P.

Similarly, the potential V_2 at P due to q_2 and V_3 due to q_3 are given by

$$V_2 = \frac{1}{4\pi\varepsilon_0}\frac{q_2}{r_{2p}}, V_3 = \frac{1}{4\pi\varepsilon_0}\frac{q_3}{r_{3p}}$$

Where r_{2p} and r_{3p} are the distances of P from charges q_2 and q_3, respectively; and so on for the potential due to other charges.

By the superposition principle, the potential V at P due to the total charge configuration is the algebraic sum of the potentials due to the individual charges

$$V = V_1 + V_2 + \ldots\ldots\ldots + V_n$$

$$= \frac{1}{4\pi\varepsilon_0}\left(\frac{q_1}{r_{1p}} + \frac{q_2}{r_{2p}} + \cdots \ldots \ldots + \frac{q_n}{r_{np}}\right) \qquad (5.18)$$

If we have a continuous charge distribution characterised by a charge density $\rho(\boldsymbol{r})$, we divide it, as before , into small volume elements each of size Δv and carrying a charge $\rho\Delta v$. we then determine the potential due to each volume element and sum (strictly speaking, integrate) over all such contributions, and thus determine the potential due to the entire distribution.

We have seen in Chapter 1 that for a uniformly charges spherical shell, the electric field outside the shell is as if the entire charge is concentrated at the centre. Thus, the potential outside the shell is given by

$$V = \frac{1}{4\pi\varepsilon_0}\frac{q}{r} \; (r \geq R) \qquad [5.19\,(a)]$$

Where q is the total charge on the shell and R its radius. The electric field inside the shell is zero. This implies (Section 5.6) that potential is constant inside the shell (as no work is done in moving a charge inside the shell), and therefore, equals its value at the surface , which is

$$V = \frac{1}{4\pi\varepsilon_0}\frac{q}{R} \qquad [5.19\,(b)]$$

5.6 EQUIPOTENTIAL SURFACES

An equi potential surface is a surface with a constant value of potential at all points on the surface. For a single charge q. The potential is given by Eq.(5.8):

$$V = \frac{1}{4\pi\varepsilon_0}\frac{q}{r}$$

This shows that V is a constant if r is constant. Thus, equipotential surface of a single point charge are concentric spherical surfaces centred at the charge.

Now the electric field lines for a single charge q are radial lines starting from or ending at the charge, depending on whether q is positive or negative. Clearly, the electric field at every point is normal to the equipotential surface passing through that point. This is true in general : *for any charge configuration. Equipotential surface through a point is normal to the electric field at that point.* The proof of this statement is simple.

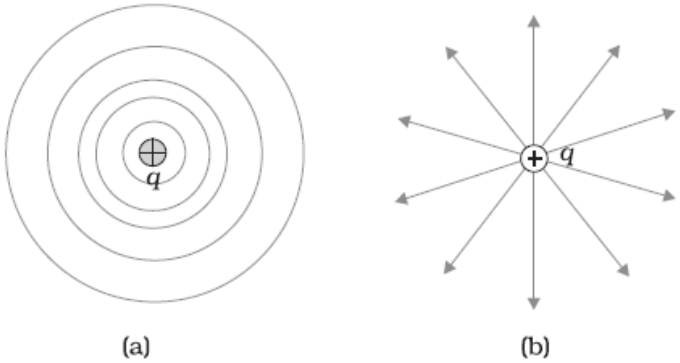


Fig. 5.8 For a single charge *q*(a) equipotential surfaces are spherical surfaces centred at the charge, and (b) electric field lines are radial, starting from the charge if *q* > 0.

If the field were not normal to the equipotential surface, it would have non-zero component along the surface. To move a unit test charge against the direction of the component of the field, work would have to be done. But this is in contradiction to the definition of an equipotential surface: there is no potential difference between any two points on the surface and no work is required to move a test charge on the surface. The electric field must, therefore, be normal to the equipotential surface at every point. Equipotential surface offer an alternative visual picture in addition to the picture of electric field lines around a charge configuration.

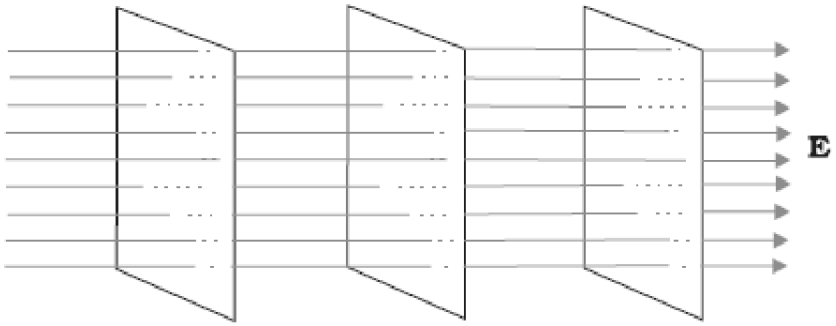


Fig. 5.9 Equipotential surfaces for a uniform electric field

For a uniform electric field **E,** say, along the x-axis, the equipotential surfaces are planes normal to the x-axis , i.e., planes parallel to the y-z plane (Fig. 5.9). Equipotential surfaces for (a) a dipole and (b) two identical positive charges are shown in Fig . 5.10

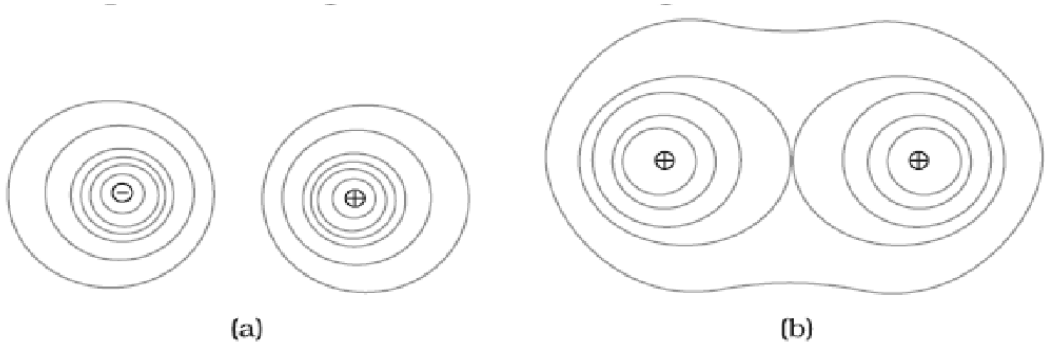


Fig. 5.10 Some equipotential surfaces for (a) a dipole, (b) two identical positive charges.

5.6.1 Relation between field and potential

Consider two closely spaced equipotential A and B (Fig.5.11) with potential values V and V+ δV, where δV is the change in V in the direction of the electric field $\boldsymbol{E}$. Let P be a point on the surface A from P. Imagine that a unit positive charge is moved along this perpendicular from the surface B to surface A against the electric field.

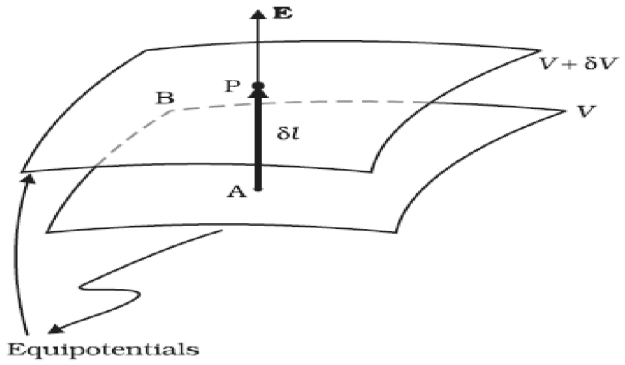


Fig. 5.11 From the potential to the field.

The work done in this process is $|\boldsymbol{E}|\ \delta l$. This work equals the potential difference $V_A - V_B$.

Thus,

$$|\boldsymbol{E}|\delta l = V - (V + \delta V) = -\delta V$$

i.e., $$|\boldsymbol{E}| = \frac{\delta V}{\delta l} \qquad (5.20)$$

Since δV is negative, δV = - $|\delta V|$, we can rewrite Eq (5.20) as

$$|\boldsymbol{E}| = -\frac{\delta V}{\delta l} = +\frac{|\delta V|}{\delta l} \qquad (5.21)$$

We thus arrive at two important conclusions concerning the relation between electric field and potential:

(i) *Electric field is in the direction in which the potential decreases steepest.*

(ii) *Its magnitude is given by the charge in the magnitude of potential per unit displacement normal to the equipotential surface at the point*

5.7 POTENTIAL ENERGY OF A SYSTEM OF CHARGES

Consider first the simple case of two charges q_1 and q_2 with position vector $\mathbf{r}_1$ and $\mathbf{r}_2$ relative to some origin. Let us calculate the work done (externally) in building up this configuration. This means that we consider the charges q_1 and q_2 initially at infinity and determine the work done by an external agency to bring the charges to the given locations. Suppose, first the charge q_1 is brought from infinity to the point r_1.There is no external field against which work needs to be done, so work done in bringing q_1 from infinity to r_1 is zero. This charge produces a potential in space given by

$$V_1 = \frac{1}{4\pi\varepsilon_0}\frac{q_1}{r_{1p}}$$

Where r_{1p} is the distance of a point P in space from the location of q_1. From the definition of potential, work done in bringing charge q_2 from infinity to the point $\mathbf{r}_2$ is q_2 times the potential at $\mathbf{r}_2$ due to q_1 :

$$\text{Work done on } q_2 = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r_{12}}$$

Where r_{12} is the distance between points 1 and 2.

Since electrostatic force is conservative, this work gets stored in the form of potential energy of the system. Thus, the potential energy of a system of two charges q_1 and q_2 is

$$U = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r_{12}} \qquad (5.22)$$

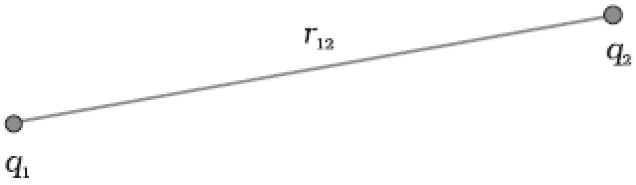


Fig. 5.12 Potential energy of a system of charges q_1 and q_2 is directly proportional to the product of charges and inversely to the distance between them.

Obviously, if q_2 was brought first to its present location and q_1 brought later, the potential energy U would be the same. Eq.(5.22) , is unaltered whatever way the charges are brought to the specified locations, because of path-independence of work for electrostatic force.

Equation (5.22) is true for any sign of q_1 and q_2. If $q_1q_2 > 0$, potential energy is positive. This is an expected, since for like charge [$q_1q_2 < 0$], the electrostatic force is attractive. In that case, a positive amount of work is needed against this force to take the charges from the given location to infinity. In other words, a negative amount of work is needed for the reverse path (from infinity to the present locations), so the potential energy is negative.

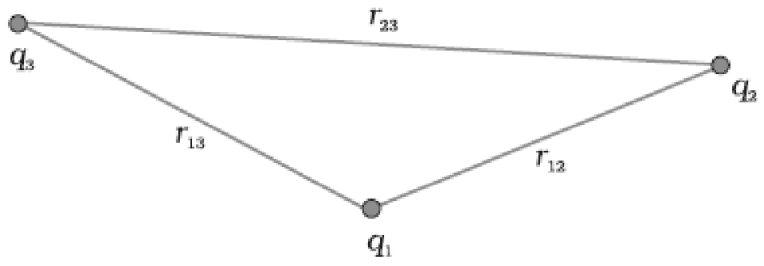


Fig. 5.13 Potential energy of a system of three charges is given by Eq. (2.26), with the notation given in the figure.

Equation (5.22) is easily generalised for a system of any number of point charges. Let us calculate the potential energy of a system of three charges q_1, q_2 and q_3 located at $\mathbf{r}_1$, $\mathbf{r}_2$, $\mathbf{r}_3$, respectively . To bring q_1 first from infinity to $\mathbf{r}_1$, no work is required. Next we bring q_2 from infinity to $\mathbf{r}_2$. As before, work done in this step is

$$q_2 V_1(\boldsymbol{r}_2) = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r_{12}} \quad (5.23)$$

The charges q_1 and q_2 produce a potential, which at any point P is given by

$$V_{1.2} = \frac{1}{4\pi\varepsilon_0}\left(\frac{q_1}{r_{1p}} + \frac{q_2}{r_{2p}}\right) \quad (5.24)$$

Work done next in bringing q_3 from infinity to the point $\mathbf{r}_3$ is q_3 times $V_{1.2}$ at $\mathbf{r}_3$.

$$q_3 V_{1.2}(\boldsymbol{r}_3) = \frac{1}{4\pi\varepsilon_0}\left(\frac{q_1 q_3}{r_{13}} + \frac{q_2 q_3}{r_{23}}\right) \quad (5.25)$$

The total work done in assembling the charges at the given locations is obtained by adding the work done in different steps [Eq.(5.23) and Eq.(5.25)],

$$U = \frac{1}{4\pi\varepsilon_0}\left(\frac{q_1 q_2}{r_{12}} + \frac{q_1 q_3}{r_{13}} + \frac{q_2 q_3}{r_{23}}\right) \quad (5.26)$$

Again, because of the conservative nature of the electrostatic force [or equivalently , the path independence of work done], the final expression for U, Eq. (5.26), is independent of the manner in which the configuration is assembled. *The potential energy is characteristic of the present state of configuration, and not the way the state is achieved.*

5.8 POTENTIAL ENERGY IN AN EXTERNAL FIELD

5.8.1 Potential energy of a single charge

In Section 5.7, the source of the electric field was specified-the charges and their locations – and the potential energy of the system of those charges was determined. In this section, we ask a related but a distinct question. What is the potential energy of a charge q in a given field? This question was, in fact, the starting point that led us to the notion of the electrostatic potential (Sections 5.1 and 5.2). But here we address this question again to clarify in what way it is different from the discussion in Section 5.7.

The main difference is that we are now concerned with the potential energy of a charge (or charges) in an *external* field. The external field **E** ia not produced by the given charge(s) whose potential energy we wish to calculate. **E** is produced by sources external to the given charge(s). The external sources may be known, but often they are unknown or unspecified; what is specified is the electric field **E** or the electrostatic potential V due to the external sources. We assume that the charge q does not significantly affect the sources producing the external field. This is true if q is very small, or the external sources are held fixed by other unspecified forces. Even if q is finite, its influence on the external sources may still be ignored in the situation when very strong sources far away at infinity produce a finite field **E** in the region of interest. Note again that we are interested in determining the potential energy of a given charge q (and later, a system of charges) in the external field ; we are not interested in the potential energy of the sources producing the external electric field.

The external electric field **E** and the corresponding external potential V may vary from point to point. By definition, V at a point P is the work done in bringing a unit positive charge from infinity to the point P.(We continue to take potential at infinity to be zero). Thus, work done in bringing a charge q from infinity to the point P in the external field is q V. This work is stored in the form of potential energy of q. If the point P has position vector **r** relative to some origin, we can write:

Potential energy of q at **r** in an external field

$$= qV(\mathbf{r}) \quad (5.27)$$

Where V(**r**) is the external potential at the point **r .**

Thus, if an electron with charge q = e = 1.6 x 10^{-19} C is accelerated by a potential difference of ΔV = 1 volt, it would gain energy of qΔV = 1.6 x 10^{-19} J . This unit of energy is defined as 1 *electron volt* or 1 e V , i.e., 1 eV = 1.6 x 10^{-19} J. The units based on eV are most commonly used in atomic, nuclear and particle physics, (1 keV = 10^3eV = 1.6 x 10^{-16} J, 1 MeV = 10^6eV = 1.6 x 10^{-13}J, 1 GeV = 10^9 eV = 1.6 x 10^{-10} J and 1 TeV = 10^{12} eV = 1.6 x 10^{-7} J).

5.8.2 Potential energy of a system of two charges in an external field

Next, we ask: what is the potential energy of a system of two charges q_1 and q_2 located at $\mathbf{r}_1$ and $\mathbf{r}_2$ respectively, in an external field? First, we calculate the work done in bringing the charge q_1 from infinity to $\mathbf{r}_1$. Work done in this step is q_1 V($\mathbf{r}_1$) , using Eq.(5.27). Next, we consider the work done in bringing q_2 to $\mathbf{r}_2$. In this step, work is done not only against the external field **E** but also against the field due to q_1.

Work done on q_2 against the external field

$$= q_2 \, V(\mathbf{r}_2)$$

Work done on q_2 against the field due to q_1

$$= \frac{q_1 q_2}{4\pi\varepsilon_0 r_{12}}$$

Where r_{12} is the distance between q_1 and q_2. We have made use of Eqs. (5.27) and (5.22). By the superposition principle for fields, we add up the work done on q_2 against the two fields (**E** and that due to q_1):

Work done in bringing q_2 to $\mathbf{r}_2$

$$= q_2 \, V(\mathbf{r}_2) + \frac{q_1 q_2}{4\pi\varepsilon_0 r_{12}} \quad (5.28)$$

Thus, Potential energy of the system

= the total work done in assembling the configuration

$$= q_1 \, V(\mathbf{r}_1) + q_2 \, V(\mathbf{r}_2) + \frac{q_1 q_2}{4\pi\varepsilon_0 r_{12}} \quad (5.29)$$

5.8.3 Potential energy of a dipole in an external field

Consider a dipole with charges q_1 = +q and q_2 = - q placed in a uniform electric field **E**, as shown in Fig. 5.14.

As seen in the last chapter , in a uniform electric field, the dipole experiences no net force; but experiences a torque τ given by

$$\tau = p \times E \quad (5.30)$$

which will tend to rotate it (unless **p** is parallel or antiparallel to **E**).

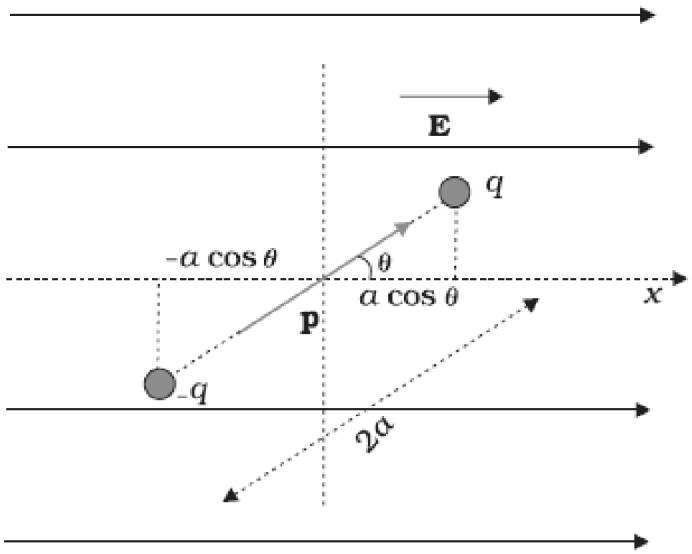


Fig. 5.14 Potential energy of a dipole in a uniform external field.

Suppose an external torque τ_{ext} is applied in such a manner that it just neutralises this torque and rotates it in the plane of paper from angle θ_0 to angle θ_1 at an infinitesimal angular speed and without angular acceleration. The amount of work done by the external torque will be given by

$$W = \int_{\theta_0}^{\theta_1} \tau_{ext}(\theta) d\theta = \int_{\theta_0}^{\theta_1} pE \sin\theta \, d\theta$$

$$= pE(cos\theta_0 - \cos\theta_1) \qquad (5.31)$$

This work is stored as the potential energy of the system. We can then associate potential energy U(θ) with an inclination θ of the dipole. Similar to other potential energies, there is a freedom in choosing the angle where the potential energy U is taken to be zero. A natural choice is to take θ_0 = /2 . (An explanation for it is provided towards the end of discussion.) We can then write.

$$U(\theta) = pE\left(\cos\frac{\pi}{2} - \cos\theta\right) = pE\cos\theta = -\mathbf{p}.\mathbf{E} \qquad (5.32)$$

This expression can alternately be understood also from Eq.(5.29). We apply Eq.(5.29) to the present system of two charges +q and –q . The potential energy expression then reads

$$U'(\theta) = q[V(\boldsymbol{r}_1) - V(\boldsymbol{r}_2)] - \frac{q^2}{4\pi\varepsilon_0 \times 2a} \qquad (5.33)$$

Here , $\mathbf{r}_1$ and $\mathbf{r}_2$ denote the position vectors of +q and –q. Now , the potential difference between positions $\mathbf{r}_1$ and $\mathbf{r}_2$ equals the work done in bringing a unit positive charge against field from $\mathbf{r}_2$ to $\mathbf{r}_1$. The displacement parallel to the force is 2a cosθ . Thus, [V($\mathbf{r}_1$)- V($\mathbf{r}_2$)] =

$$U'(\theta) = -pE\,cos\theta - \frac{q^2}{4\pi\varepsilon_0 \times 2a} = -p.E - \frac{q^2}{4\pi\varepsilon_0 \times 2a} \qquad (5.34)$$

We note that $U'(\theta)$ differs from $U(\theta)$ by a quantity which is just a constant for a given dipole. Since a constant is insignificant for potential energy, we can drop the second term in Eq.(5.34) and it then reduces to Eq.(5.32).

We can now understand why we took $\theta_2 = \pi/2$. In this case, the work done against the *external* field E in bringing +q and –q are equal and opposite and cancel out, i.e., q[V($\mathbf{r}_1$)-V($\mathbf{r}_2$)]=0.

5.9 ELECTROSTATICS OF CONDUCTORS

Conductors and insulators were described briefly in Chapter 4. Conductors contain mobile charge carriers. In metallic conductors, these charge carriers are electrons. In a metal, the outer (valence) electrons part away from their atoms and are free to move. These electrons are free within the metal but not free to leave the metal. The free electrons form a kind of 'gas'; they collide with each other and with the ions, and move randomly in different directions. In an external electric field, they drift against the direction of the field. The positive ions made up of the nuclei and the bound electrons remain held in their fixed positions. In electrolytic conductors, the charge carries are both positive and negative ions; but the situation in this case is more involved – the movement of the charge carries is affected both by the external electric field as also by the so-called chemical forces (see Chapter 6). We shall restrict our discussion to metallic solid conductors. Let us note important results regarding electrostatics of conductors.

1. inside a conductor, electrostatic field is zero

Consider a conductor, neutral or charged. There may also be an external electrostatic field. In the static situation, when there is no current inside or on the surface of the conductor, the electric field is zero everywhere inside the conductor. This fact can be taken as the defining property of a conductor. A conductor has free electrons. As long as electric field is not zero, the free charge carries would experience force and drift. In the static situation, the free charges have so distributed themselves that the electric field is zero everywhere inside. *Electrostatic field is zero inside a conductor.*

2. At the surface of a charged conductor, electrostatic field must be normal to the surface at every point

If **E** were not normal to the surface, it would have some non-zero component along the surface. Free charges on the surface of the conductor would then experience force and move. In the static situation, therefore, **E** should have no tangential component. Thus *electrostatic field at the surface of a charged conductor must be normal to the surface at every point.* (For a conductor without any surface charge density, field is zero even at the surface.) See result 5.

3. The interior of a conductor can have no excess charge in the static situation

A neutral conductor has equal amounts of positive and negative charges in every small volume or surface element. When the conductor is charged, the excess charge can reside only on the surface in the static situation. This follows from the Gauss's law. Consider any arbitrary volume element, electrostatic field is zero. Thus the total electric flux through S is zero. Hence, by Gauss's law, there is no net charge enclosed by S. But the surface S can be made as small as you like, i.e., the volume v can be made vanishingly small. This means *there is no net charge at any point inside the conductor, and any excess charge must reside at the surface.*

4. Electrostatic potential is constant throughout the volume of the conductor and has the same value (as inside) on its surface

This follows from results 1 and 2 above. Since E = 0 inside the conductor and has no tangential component on the surface, no work is done in moving a small test charge within the conductor and on its surface. That is, there is no potential difference between any two points inside or on the surface of the conductor. Hence, the result, If the conductor is charged, electric field normal to the surface exists. This means potential will be different for the surface and a point just outside the surface.

In a system of conductors of arbitrary size, shape and charge configuration, each conductor is characterised by a constant value of potential, but this constant may differ from one conductor to the other.

5. Electric field at the surface of a charged conductor

$$\mathbf{E} = \frac{\sigma}{\varepsilon_0}\hat{\boldsymbol{n}} \qquad (5.35)$$

Where $\sigma$the surface is charge density and $\hat{\boldsymbol{n}}$ is a unit vector normal to the surface in the outward direction.

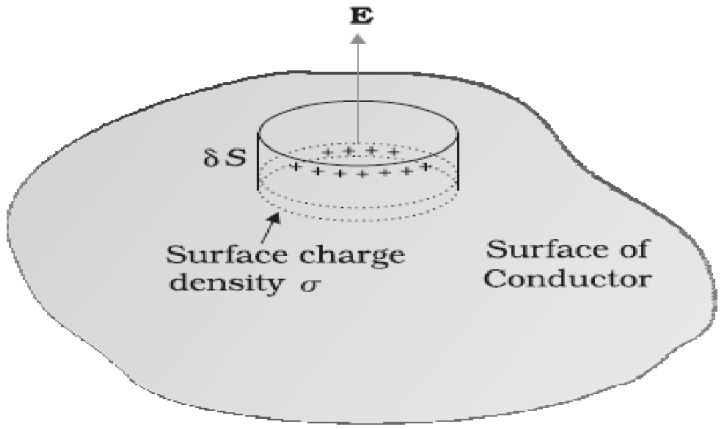


Fig. 5.15 The Gaussian surface (a pill box) chosen to derive Eq. (2.35) for electric field at the surface of a charged conductor.

To derive the result, choose a pill box (a short cylinder) as the Gaussian surface about any point P on the surface, as shown in Fig. 5.15. The pill box is partly inside and partly outside the surface of the conductor. It has a small area of cross section δS and negligible height.

Just inside the surface, the electrostatic field is zero; just outside, the field is normal to the surface with magnitude E. Thus, the contribution to the total flux through the pill b comes only from the outside (circular) cross-section of the pill box. This equals $\pm$ EδS (positive for $\sigma > 0$, negative for $\sigma < 0$), since over the small area δS are parallel or antiparallel. The charge enclosed by the pill box is $\sigma\delta S$.

By Gauss's law

$$E\delta S = \frac{|\sigma|\delta S}{\varepsilon_0}$$

$$E = \frac{|\sigma|}{\varepsilon_0} \qquad (5.36)$$

Including the fast that electric field is normal to the surface, we get the vector relation, Eq. (5.35), which is true for both signs of . For $\sigma > 0$, electric field is normal to the surface outward; for $\sigma < 0$, electric field is normal to the surface inward.

6. Electrostatic shielding

Consider a conductor with a cavity, with no changes inside the cavity. A remarkable result is that the electric field inside the cavity is zero, whatever be the size and shape of the cavity and whatever be the charge on the conductor and the external fields in which it might be placed. We have proved a simple case of this result already; the electric field inside a charged spherical shell is zero. The proof of the result for the shall makes use of the spherical symmetry of the shell (see Chapter 4).

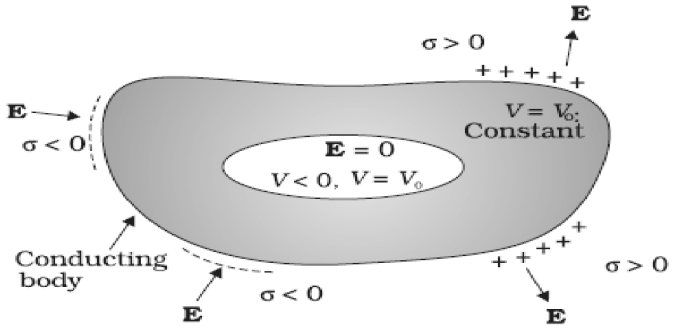


Fig. 5.16 The electric field inside a cavity of any conductor is zero. All charges reside only on the outer surface of a conductor with cavity. (There are no charges placed in the cavity.)

But the vanishing of electric field in the (charge-free) cavity of a conductor is, as mentioned above, a very general result. A related result is that even if the conductor is charged or charges are induced on a neutral conductor by an external field, all charges reside only on the outer surface of a conductor with cavity.

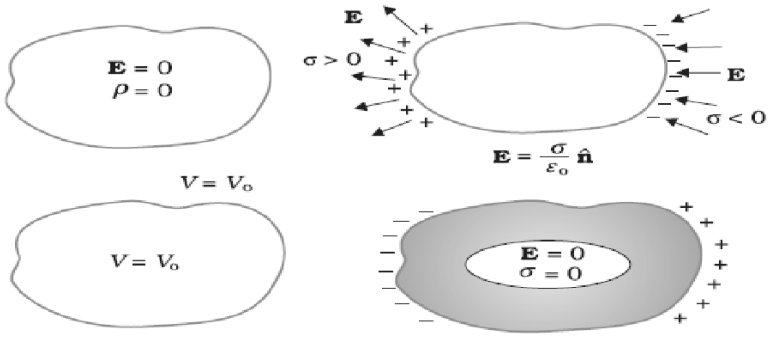


Fig. 5.17 Some important electrostatic properties of a conductor.

The proofs of the results noted in Fig.5.16 are omitted here, but we note their important implication. Whatever be the charge and field configuration outside, any cavity in a conductor remains shielded from outside electric influence; *the field inside the cavity is always zero*. This is known as *electrostatic shielding*. The effect can be made use of in protecting sensitive instruments from outside electrical influence. Fig. 5.17 gives a summary of the important electrostatic properties of a conductor.

5.10 DIELECTRICS AND POLARISATION

Dielectrics are non-conducting substances. In contrast to conductors, they have no (or negligible number of) charge carries. Recall from Section 5.9 what happens when a conductor is placed in an external electric field. The free charge carries move and charge distribution in the conductor adjusts itself in such a way that the electric field due to induced charges opposes the external field within the conductor. This happens until, in the static situation, the two fields cancel each other and the net electrostatic field in the conductor is zero. In a dielectric, this free movement of charges is not possible. It turns out

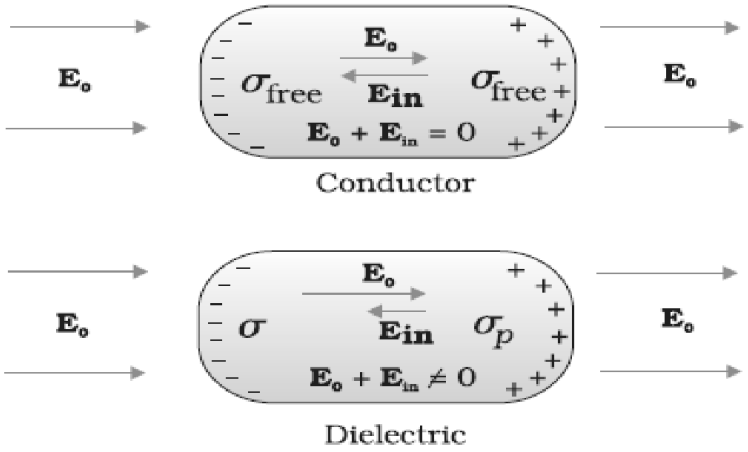


Fig. 5.18 Difference in behavior of a conductor and a dielectric in an external electric field.

that the external field induces dipole moment by stretching or re-orienting molecular dipole moments is net charges onthe surface of the dielectric which produces a field that opposes the external field. Unlike in a conductor, however, the opposing field so induced does not exactly cancel the external field. It only reduces it. The extent of the effect depends on the nature of the dielectric. To understand the effect, we need to look at the charge distribution of a dielectric at the molecular level.

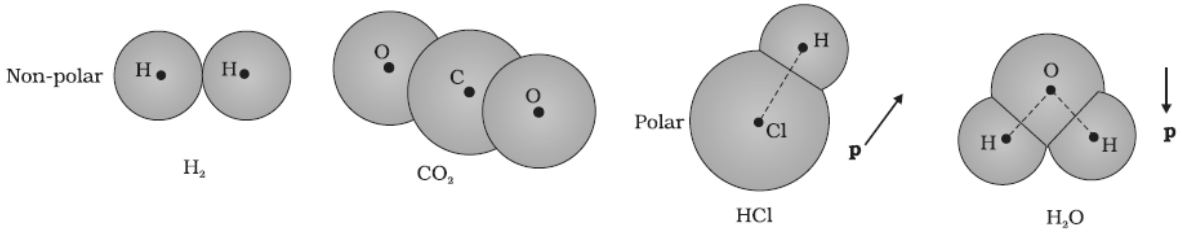


Fig. 5.19 Some examples of polar and non-polar molecules.

The molecules of a substance may be polar or non-polar. In a non-polar molecule, the centres of positive and negative charges coincide. The molecule then has no permanent (or intrinsic) dipole moment. Examples of non-polar molecules are oxygen (O_2) and hydrogen (H_2) molecules which, because of their symmetry, have no dipole moment. On the other hand, a polar molecules is one in which the centres of positive and negative charges are separated (even when there is no external field). Such molecules have a permanent dipole moment. An ionic molecule such as HCl or molecules of water (H_2O) are examples of polar molecules.

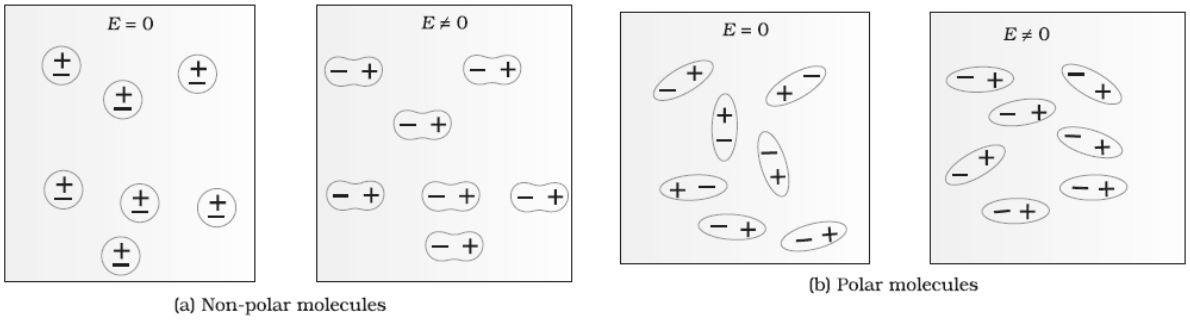


Fig. 5.20 A dielectric develops a net dipole moment in an external electric field. (a) Non-polar molecules, (b) Polar molecules.

In an external electric field, the positive and negative charges of a non-polar molecule are displaced in opposite directions. The displacement stops when the external force on the constituent charges of the molecule is balanced by the restoring force (due to internal fields in the molecule.) The non-polar molecule thus develops an induced dipole moment. The dielectric is said to be polarised by the external field. We consider only the simple situation when the induced dipole moment is in the direction of the field and is proportional to the field strength. (Substances for which this assumption is true are called linear isotropic dielectrics). The induced dipole moments of different molecules ass up giving a net dipole moment of the dielectric in the presence of the external field.

A dielectric with polar molecules also develops a net dipole moment in an external field, but for a different reason. .In the absence of any external field, the different permanent dipoles are oriented randomly due to thermal agitation; so the total dipole moment is zero. When an external field is applied, the individual dipole moments tend to align with the field. When summed overall the molecules, there is then a net dipole moment in the direction of the external field, i.e., the dielectric is polarised. The extent of polarisation depends on the relative strength of two mutually opposite factors; the dipole potential energy in the external field tending to align the dipoles with the field and thermal energy tending to disrupt the alignment. There may be, in addition, the 'induced dipole moment' effect as for non-polar molecules, but generally the alignment effect is more important for polar molecules.

Thus in either case, whether polar or non-polar, a dielectric develops a net dipole moment in the presence of an external field. The dipole moment per unit volume is called *polarisation* and is denoted by **P**. For linear isotropic dielectrics,

$$P = X_e E \qquad (5.37)$$

where X_e is a constant characteristic of the dielectric and is known as the *electric susceptibility* of the dielectric medium.

It is possible to relate X_e to the molecular properties of the substance, but we shall not pursue that here. The question is: how does the polarised dielectric modify the original external field inside it? Let us consider, for simplicity, rectangular dielectric slab placed in a uniform external field $\mathbf{E_0}$ parallel to two of its faces. The field causes a uniform polarisation **P of** the dielectric. Thus every volume element Δv of the slab has a dipole moment $\mathbf{P}\Delta\boldsymbol{v}$ in the direction of the field. The volume element Δv is macroscopically small but contains a very large number of molecular dipole. Anywhere inside the dielectric, the volume element Δv has no net charge (though it has net dipole moment). This is, because, the positive charge of one dipole sits close to the negative charge of the adjacent dipole. However, at the surfaces of the dielectric normal to the electric field, there is evidently a net charge density. As seen in Fig 5.21, the positive ends of the dipoles remain un-neutralised at the right surface and the negative ends at the left surface. The unbalanced charges are the induced charges due to the external field.

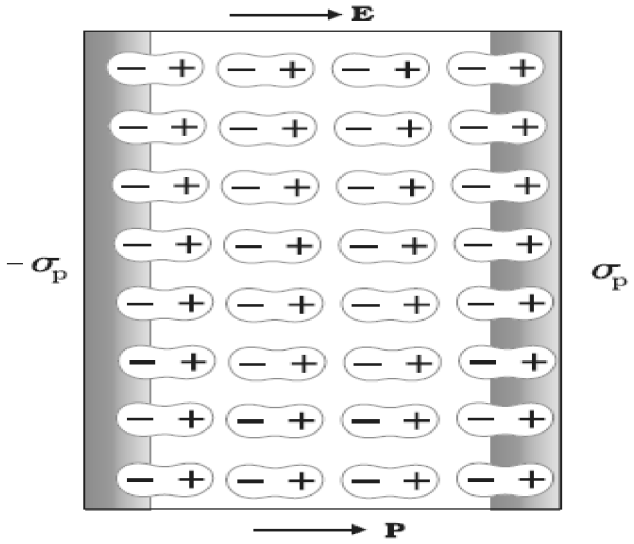


Fig. 5.21 A uniformly polarised dielectric amounts to induced surface charge density, but no volume charge density.

Thus, the polarised dielectric is equivalent to two charged surface with induced surface charge densities, say σ_p and - σ_p. Clearly, the field produced by these surface charges opposes the external field. The total field in the dielectric is, thereby, reduced from the case when no dielectric is present. We should note that the surface charge density $\pm\sigma_p$ arises from bound (not free charges) in the dielectric.

5.11 CAPACITORS AND CAPACITANCE

A capacitor is a system of two conductors separated by an insulator (Fig.5.22). The conductors have charges, say Q_1 and Q_2, and potentials V_1 and V_2. Usually, in practice, the two conductors have charges Q and – Q, with potential difference $V = V_1 – V_2$ between them. We shall consider only this kind of charge configuration of the capacitor. (Ever a single conductor can be used as a capacitor by assuming the other at infinity). The conductors may be so charged by connecting them to the two terminals of a battery. Q is called the charge of

the capacitor, though this, in fact, is the charge on one of the conductors – the total charge of the capacitor is zero.

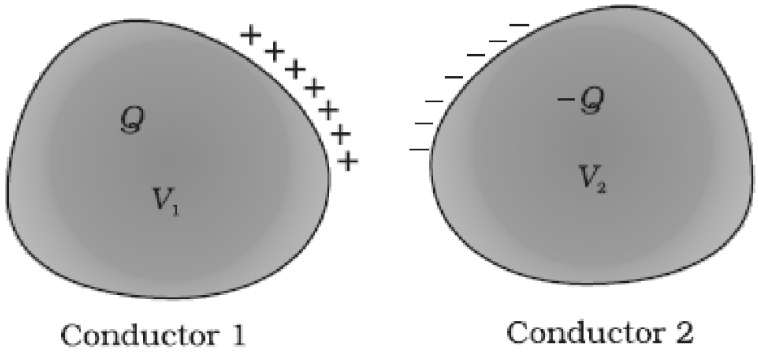


Fig. 5.22 A system of two conductors separated by an insulator forms a capacitor.

The electric field in the region between the conductors is proportional to the charge Q. That is, if the charge on the capacitor is, say doubled, the electric field will also be doubled at every point. (This follows from the direct proportionality between field and charge implied by Coulomb's law and the superposition principle). Now, potential difference V is the work done per unit positive charge in taking a small test charge from the conductor 2 to 1 against the field. Consequently, V is also proportional to Q, and the ratio Q/V is a constant:

$$C = \frac{Q}{V} \qquad (5.38)$$

The constant C is called the *capacitance* of the capacitor. C is independent of Q or V, as stated above. The capacitance C depends only on the geometrical configuration (shape, size, separation) of the system of two conductors. [As we shall see later, it also depends on the nature of the insulator (dielectric) separating the two conductors]. The SI unit of capacitance is 1 farad (= 1 coulomb volt^{-1}) or 1 F = 1 C V^{-1}. A capacitor with fixed capacitance is symbolically shown as, ⊣⊢ while the one with variable capacitance is shown as ⫤⫥

Equation (5.38) shows that for large C, V is small for a given Q. This means a capacitor with large capacitance can hold large amount of charge Q at a relatively small V. This is of practical importance. High potential difference implies strong electric field around the conductors. A strong electric field can ionise the surrounding air and accelerate the charges so produced to the oppositely charged plates, thereby neutralising the charge on the capacitor plates, at least partly. In other words, the charge of the capacitor leaks away due to the reduction in insulating power of the intervening medium.

The maximum electric field that a dielectric medium can withstand without break-down (of its insulating property) is called its *dielectric strength*; for it is about 3 x 10^6 Vm^{-1} . For a separation between conductors of the order of 1 cm or so, this field corresponds to a potential difference of 3 x 10^4 V between the conductors. Thus, for a capacitor to store a large amount of charge without leaking , its capacitance should be high enough so that the potential difference and hence the electric field do not exceed the break-down limits. Put differently , there is a limit to the amount of charge that can be stored on a given capacitor without significant leaking. In practice, a farad is a very big unit ; the most common units are its sub-multiples 1 $\mu F = 10^{-6} F$, 1 $nF = 10^{-9} F$, 1 $pF = 10^{-12} F$, etc. Besides its use in storing charge, a capacitor is a key element of most ac circuits with important functions, as described in chapter 10.

5.12 THE PARALLEL PLATE CAPACITOR

A parallel plate capacitor consists of two large plane parallel conducting plates separated by a small distance (Fig .5.23).

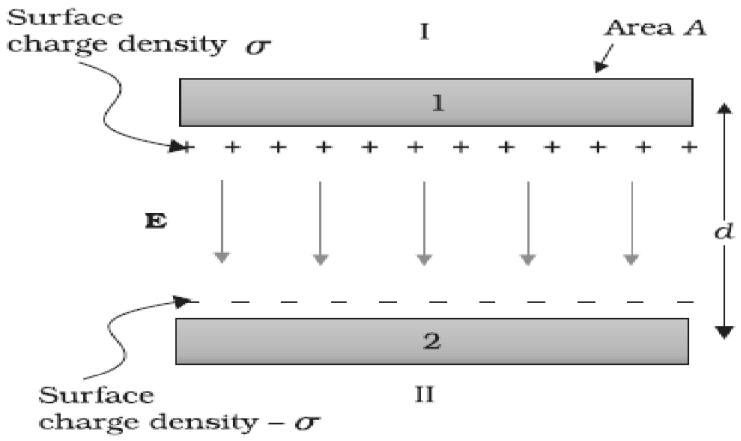


Fig. 5.23 The parallel plate capacitor.

We first take the intervening medium between the plates to be vacuum. The effect of a dielectric medium between the plates in discussed in the next section. Let A be the area of each plate and d the separation between them. The two plates have charges Q and – Q. Since d is much smaller than the linear dimension of the plates ($d^2 << A$), we can use the result on electric field by an infinite plate sheet of uniform surface charge density. Plate 1 has surface charge density $\sigma = Q/A$ and plate 2 has a surface charge density $-\sigma$. Using Eq. (5.33), the electric field in different regions is:

Outer region 1 (region above the plate 1),

$$E = \frac{\sigma}{2\varepsilon_0} - \frac{\sigma}{2\varepsilon_0} = 0 \qquad (5.39)$$

Outer region II (region below the plate 2).

$$E = \frac{\sigma}{2\varepsilon_0} - \frac{\sigma}{2\varepsilon_0} = 0 \qquad (5.40)$$

In the inner region between the plates 1 and 2, the electric fields due to the two charged plates add up, giving

$$E = \frac{\sigma}{2\varepsilon_0} + \frac{\sigma}{2\varepsilon_0} = \frac{\sigma}{\varepsilon_0} = \frac{Q}{\varepsilon_0 A} \qquad (5.41)$$

The direction of electric field is from the positive to the negative plate.

Thus, the electric field is localised between the two plates and is uniform throughout. For plates with finite area, this will not be true near the outer boundaries of the plates. The field lines bend outward at the edges – an effect called 'fringing of the field'. By the same token, σ will not be strictly uniform on the entire plate. [E and σ are related by Eq. (5.35).] However, for $d^2 << A$, these effects can be 9ignored in the regions sufficiently far from the edges, and the field there is given by Eq.(5.41). Now for uniform electric field, potential difference is simply the electric field times the distance between the plates, that is,

$$V = Ed = \frac{1}{\varepsilon_0}\frac{Qd}{A} \qquad (5.42)$$

The capacitance C of the parallel plate capacitor is then

$$C = \frac{Q}{A} = \frac{\varepsilon_0 A}{d} \qquad (5.43)$$

which, as expected, depends only on the geometry of the system. For typical values like A = 1 m^2 , d = 1 mm, we get

$$C = \frac{8.85\times10^{-12}C^2N^{-1}m^{-2}\times1m^2}{10^{-3}m} = 8.85 \times 10^{-9}F \quad (5.44)$$

(You can check that if 1F = 1C V^{-1} = 1 C (NC^{-1} m)$^{-1}$ = 1 C^2 N^{-1} m^{-1}).

This shows that 1F is too big a unit in practice, as remarked earlier. Another way of seeing the 'bigness' of 1F is to calculate the area of the plates needed to have C = 1F for a separation of, say 1 cm:

$$A = \frac{Cd}{\varepsilon_0} = \frac{1F\times10^{-2}m}{8.85\times10^{-12}C^2N^{-1}m^{-2}} = 10^9m^2 \quad (5.45)$$

which is a plate about 30 km in length and breadth !

5.13 EFFECT OF DIELECTRIC ON CAPACITANCE

With the understanding of the behaviour of the behaviour of dielectrics in an external field developed in Section 5.10, let us see how the capacitance of a parallel plate capacitor is modified when a dielectric is present. As before, we have two large plates, each of area A, separated by a distance d. The charge on the plates is $\pm Q$, corresponding to the charge density $\pm\sigma$ (with σ = Q / A). When there is vacuum between the plates,

$$E_0 = \frac{\sigma}{\varepsilon_0}$$

and the potential difference V_0 is

$$V_0 = E_0 d$$

The capacitance C_0 in this case is

$$C_0 = \frac{Q}{V_0} = \varepsilon_0 \frac{A}{d} \quad (5.46)$$

Consider next a dielectric inserted between the plates fully occupying the intervening region. The dielectric is polarised by the field and, as explained in Section 5.10, the effect is equivalent to two charged sheets (at the surfaces of the dielectric normal to the field) with surface charge densities$\sigma_p and -\sigma_p$. The electric field in the dielectric then corresponds to the case when the net surface charge density on the plates is $\pm(\sigma - \sigma_p)$. That is,

$$E = \frac{\sigma-\sigma_p}{\varepsilon_0} \quad (5.47)$$

So that the potential difference across the plates is

$$V = Ed = \frac{\sigma-\sigma_p}{\varepsilon_0} d \quad (5.48)$$

For linear dielectrics, we expect σ_p to be proportional to E_0 , i.e., to σ,

Thus, $(\sigma - \sigma_p)$ is propoetional to σ and we can write

$$\sigma - \sigma_p = \frac{\sigma}{K} \quad (5.49)$$

where K is a constant characteristic of the dielectric. Clearly , K > 1 . We then have

$$V = \frac{\sigma d}{\varepsilon_0 K} = \frac{Qd}{A\varepsilon_0 K} \quad (5.50)$$

The capacitance C, with dielectric between the plates, is then

$$C = \frac{Q}{V} = \frac{\varepsilon_0 KA}{d} \quad (5.51)$$

The product $\varepsilon_0 K$ is called the *permittivity* of the medium and is denoted by ε

$$\varepsilon = \varepsilon_0 K \quad (5.52)$$

For vacuum K = 1 and $\varepsilon = \varepsilon_0$; ε_0 is called the *permittivity of the vacuum*. The dimensionless ratio

$$K = \frac{\varepsilon}{\varepsilon_0} \tag{5.53}$$

is called the *dielectric* constant of the substance. As remarked before, from Eq.(5.49), it is clear that K is greater than 1. From Eqs.(5.46) and (5.51)

$$K = \frac{C}{C_0} \tag{5.54}$$

Thus, the dielectric constant of a substance is the factor (>1) by which the capacitance increases from its vacuum value, when the dielectric is inserted fully between the plates of a capacitor. Though we arrived at Eq.(5.54) for the case of a parallel plate capacitor, it holds good for any type of capacitor and can , in fact, be viewed in general as a definition of the dielectric constant of a substance.

5.14 COMBINATION OF CAPACITORS

We can combine several capacitors of capacitance C_1 , C_2 ,.....C_n to obtain a system with some effective capacitance C . The effective capacitance depends on the way the individual capacitors are combined. Two simple possibilities are discussed below.

5.14.1 Capacitors in series

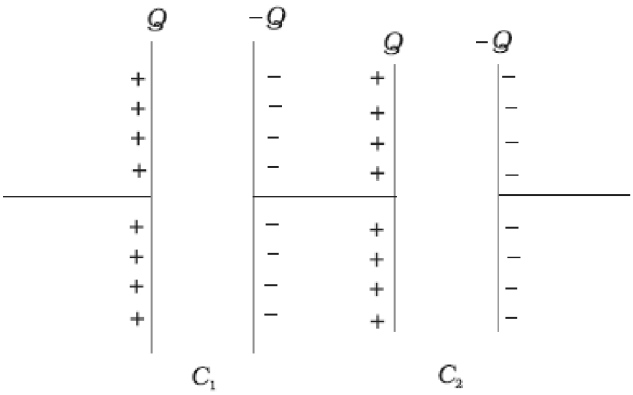


Fig. 5.24 Combination of two capacitors in series.

Fig. 5.24 shows capacitor C_1 and C_2 combined in series. The left plate of C_1 and the right plate of C_2 are connected to two terminals of a battery and have charges Q and –Q, respectively. It then follows that the right plate of C_1 has charge –Q and the left plate of C_2 has charge Q. If this was not so, the net charge on each capacitor would not be zero. This would result in an electric field in the conductor connecting C_1 and C_2 . Charge would flow until the net charge on both C_1 and C_2 is zero and there is no electric field in the conductor connecting C_1 and C_2. Thus, in the series combination, charges on the two plates $(\pm Q)$ are the same on each capacitor. The total potential drop V across the combination is the sum of the potential drops V_1 and V_2 across C_1 and C_2, respectively.

$$V = V_1 + V_2 = \frac{Q}{C_1} + \frac{Q}{C_2} \tag{5.55}$$

$$\text{i.e., } \frac{V}{Q} = \frac{1}{C_1} + \frac{1}{C_2} . \tag{5.56}$$

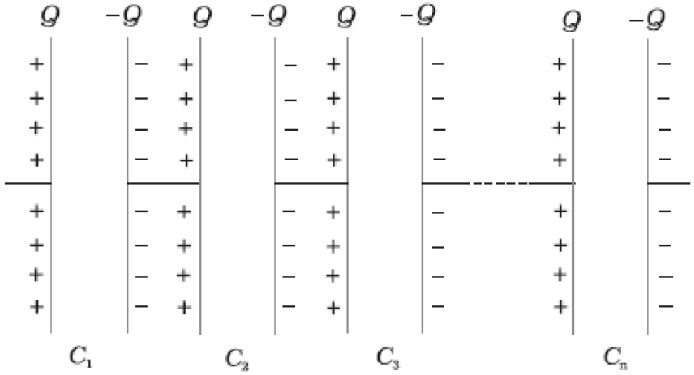


Fig. 5.25 Combination of n capacitors in series.

Now we can regard the combination as an effective capacitor with charge Q and potential difference V. The *effective capacitance* of the combination is

$$C = \frac{Q}{V} \tag{5.57}$$

We compare Eq.(5.57) with Eq.(5.56) , and obtain

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} \tag{5.58}$$

The proof clearly goes through for any number of capacitors arranged in a similar way. Equation (5.55), for n capacitors arranged in series, generalises to

$$V = V_1 + V_2 + \cdots .. + V_n = \frac{Q}{C_1} + \frac{Q}{C_2} + \cdots ... + \frac{Q}{C_n} \tag{5.59}$$

Following the same steps as for the case of two capacitors, we get the general formula for effective capacitance of a series combination of n capacitors:

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \cdots ... + \frac{1}{C_n} \tag{5.60}$$

5.14.2 Capacitors in parallel

Figure 5.26(a) shows two capacitors arranged in parallel. In this case, the same potential difference is applied across both the capacitors. But the plate charges $(\pm Q_1)$ on capacitor 1 and the plate charges $(\pm Q_2)$ on the capacitor 2 are not necessarily the same:

$$Q_1 = C_1 V, Q_2 = C_2 V \tag{5.61}$$

The equivalent capacitor is one with charge

$$Q = Q_1 + Q_2 \tag{5.62}$$

And potential difference V.

$$Q = CV = C_1 V + C_2 V \tag{5.63}$$

The effective capacitance C is, from Eq.(5.63),

$$C = C_1 + C_2 \tag{5.64}$$

The general formula for effective capacitance C for parallel combination of n capacitors [Fig. 5.26(b)] follows similarly,

$$Q = Q_1 + Q_2 + ... + Q_n \tag{5.65}$$

$$\text{i.e., } CV = C_1 V + C_2 V + ...C_n V \tag{5.66}$$

$$\text{which gives } C = C_1 + C_2 + ...C_n \tag{5.67}$$

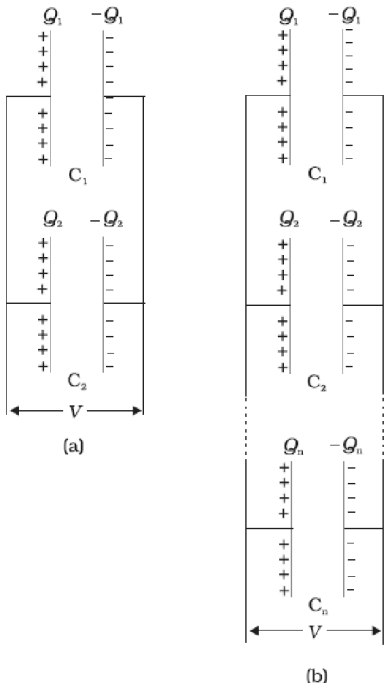


Fig. 5.26 Parallel combination of (a) two capacitors, (b) n capacitors.

5.15 ENERGY STORED IN A CAPACITOR

A capacitor, as we have seen above, is a system of two conductors with charge Q and –Q. To determine the energy stored in this configuration, consider initially two uncharged conductors 1 and 2. Imagine next a process of transferring charge from conductor 2 to conductor 1 bit by bit, so that at the end, conductor 1 gets charge Q. By charge conservation, conductor 2 has charge –Q at the end (Fig 5.27).

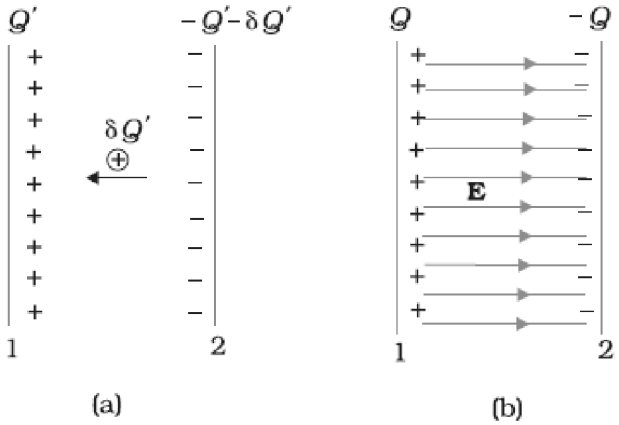


Fig. 5.27 (a) Work done in a small step of building charge on conductor 1 from Q' to $Q' + \delta Q'$. (b) Total work done in charging the capacitor may be viewed as stored in the energy of electric field between the plates.

In transferring positive charge from conductor 2 to conductor 1, work will be done externally, since at any stage conductor 1 is at a higher potential than conductor 2. To calculate the total work done, we first calculate the work done in a small step involving transfer of an infinitesimal (i.e., vanishingly small) amount of charge. Consider the intermediate situation when the conductors 1 and 2 have charges Q' and –Q' respectively. At this stage, the potential difference V' between conductors 1 to 2 is Q'/C, where C is the capacitance of the system. Next imagine that a small charge $\delta Q'$ is transferred from

conductor 2 to 1. Work done in this step ($\delta W'$), resulting in charge Q′ on conductor 1 increasing to Q'+ $\delta Q'$, is given by

$$\delta W = V'\delta Q = \frac{Q'}{C}\delta Q' \quad (5.68)$$

Since $\delta Q'$ can be made as small as we like, Eq.(5.68) can be written as

$$\delta W = \frac{1}{2C}[(Q' + \delta Q')^2 - Q'^2] \quad (5.69)$$

Equations (5.68) and (5.69) are identical because the term of second order in $\delta Q'$, i.e., $\delta Q'^2$/2C, is negligible, since $\delta Q'$ is arbitrarily small. The total work done (W) is the sum of the small work (δW) over the very large member of steps involved in building the charge Q' from zero to Q.

$$W = \sum_{sum\ over\ all\ steps} \delta W$$

$$= \sum_{sum\ over\ all\ steps} \frac{1}{2C}[(Q' + \delta Q')^2 - Q'^2] \quad (5.70)$$

$$= \frac{1}{2C}\left[\{SQ'^2 - 0\} + \{(2\delta Q')^2 - \delta Q'^2\} + \{(3\delta Q')^2 - (2\delta Q')^2\} \ldots\ldots + \{Q^2 - (Q - \delta Q)^2\}\right] \quad (5.71)$$

$$= \frac{1}{2C}[Q^2 - 0] = \frac{Q^2}{2C} \quad (5.72)$$

The same result can be obtained directly from Eq. (5.68) by integration

$$W = \int_0^Q \frac{Q'}{C}\delta Q' = \frac{1}{C}\frac{Q'^2}{2}\int_0^Q = \frac{Q^2}{2C}$$

This is not surprising since integration is nothing but summation of a large number of small terms.

We can write the final result, Eq.(5.72) in different ways

$$W = \frac{Q^2}{2C} = \frac{1}{2}CV^2 = \frac{1}{2}QV \quad (5.73)$$

Since Electrostatic force is conservative, this work is stored in the form of potential energy of the system. For the same reason, the final result for potential energy [Eq.(5.73)] is independent of the manner in which the charge configuration of the capacitor is built up. When the capacitor discharges, this stored-up energy is realised. It is possible to view the potential energy of the capacitor as 'stored' in the electric field between the plates. To see this, consider for simplicity, a parallel plate capacitor [of area A (of each plate) and separation d between the plates].

Energy stored in the capacitor

$$= \frac{1}{2}\frac{Q^2}{C} = \frac{(A\sigma)^2}{2} \times \frac{d}{\varepsilon_0 A} \quad (5.74)$$

The surface charge density σ is related to the electric field E between the plates,

$$E = \frac{\sigma}{\varepsilon_0} \quad (5.75)$$

From Eqs. (5.74) and (5.75), we get

Energy stored in the capacitor

$$U = (1/2)\ \varepsilon_0 E^2 \text{ x } Ad \quad (5.76)$$

Note that Ad is the volume of the region between the plates (where electric field alone exists). If we define *energy density as energy stored per unit volume of space*, Eq (5.76) shows that

Energy density of electric field

$$U = (1/2)\,\varepsilon_0 E^2 \qquad (5.77)$$

Though we derived Eq.(5.77) for the case of a parallel plate capacitor, the result on energy density of an electric field is, in fact, very general and holds true for electric field due to any configuration of charges.

5.16 VAN DE GRAAFF GENERATOR

This is a machine that can build up high voltages of the order of a few million volts. The resulting large electric fields are used to accelerate charged particles (electrons , protons, lons) to high energies needed for experiments to probe the small scale structure of matter. The principle underlying the machine is as follows.

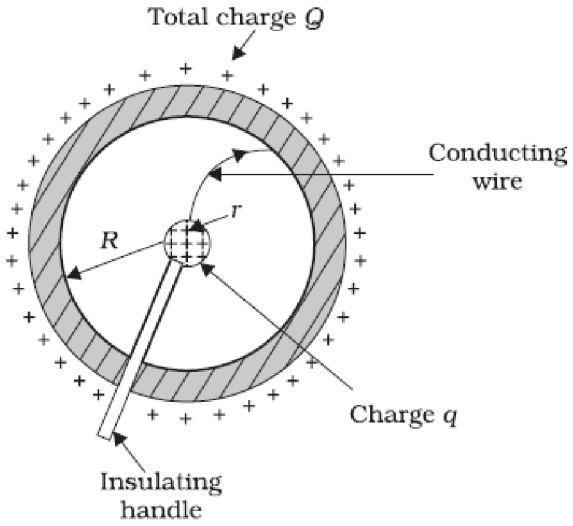


Fig. 5.28 Illustrating the principle of the electrostatic generator.

Suppose we have a large spherical conducting shell of radius R, on which we place a charge Q. This charge spreads itself uniformly all over the sphere. The field outside the sphere is just that of a point charge Q at the centre; while the field outside the sphere vanishes. So the potential outside is that of a point charge; and inside it is constant , namely the value at the radius R. We thus have:

Potential inside conducting spherical shell of radius R carrying charge Q

$$= \text{constant}$$

$$= \frac{1}{4\pi\varepsilon_0}\frac{Q}{R} \qquad (5.78)$$

Now, as shown in Fig 5.28, let us suppose that in some way we introduce a small sphere of radius r, carrying some charge q, into the large one, and place it at the centre. The potential due to this new charge clearly has the following values at radii indicated:

Potential due to small sphere of radius r carrying charge q

$$= \frac{1}{4\pi\varepsilon_0}\frac{q}{r} \text{ at surface of small sphere}$$

$$= \frac{1}{4\pi\varepsilon_0}\frac{q}{R} \text{ at large shell of radius R.} \qquad (5.79)$$

Taking both charges q and Q into account we have for the total potential V and the potential difference the values

$$V(R) = \frac{1}{4\pi\varepsilon_0}\frac{Q}{R} + \frac{q}{R}$$

$$V(r) = \frac{1}{4\pi\varepsilon_0}\frac{Q}{R} + \frac{q}{r}$$

$$V(r) - V(R) = \frac{q}{4\pi\varepsilon_0}\frac{1}{r} - \frac{1}{R} \qquad (5.80)$$

Assume now that q is positive. We see that, independent of the amount of charge Q that may have accumulated on the larger sphere and even if it is positive, the inner sphere is always at a higher potential: the difference V(r) – V(R) is positive. The potential due to Q is constant upto radius R and so cancels out in the difference !

This means that if we now connect the smaller and larger sphere by a wire, the charge q on the former will immediately flow onto the matter, even though the charge Q may be quite large. The natural tendency is for positive charge to move from higher to lower potential. Thus, provided we are somehow able to introduce the small charged sphere into the larger one, we can in this way keep piling up larger and larger amount of charge on the latter. The potential (Eq.5.78) at the outer sphere would also keep rising, at least until we reach the breakdown field of air.

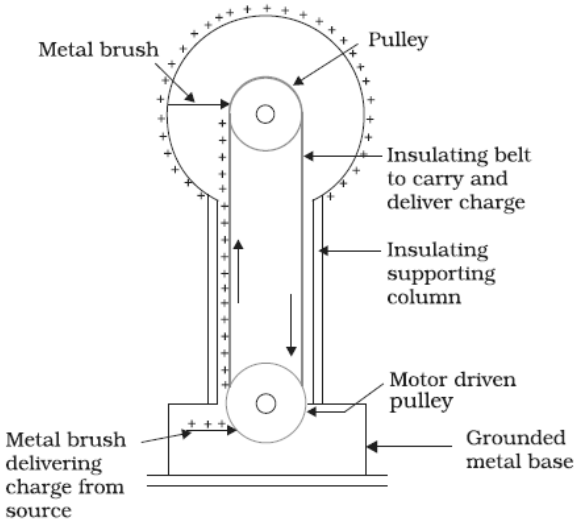


Fig. 5.29 Principle of construction of Van de Graaff generator.

This is the principle of the Van de Graaff generator. It is a machine capable of building up potential difference of a few million volts, and fields close to the breakdown field of air which is about 3 x 10^6 V/m. A schematic diagram of the Van de Graaff generator is given in Fig. 5.29. A large spherical conducting shell (of few metres radius) is supported at a height several metres above the ground on an insulating column. A long narrow endless belt insulating material, like rubber or silk, is wound around two pulley-one at ground level, one at the centre of the shell. This belt is kept continuously moving by a motor driving the lower pulley. It continuously carries positive charge, sprayed on to it by a brush at ground level, to the top. There it transfers its positive charge to another conducting brush connected to the large shell. Thus positive charge is transferred to the shell, where it spreads out uniformly on the outer surface. In this way , voltage differences of as much as 6 or 8 million volts (with respect to ground) can be built up.

SUMMARY

1. Electrostatic force is a conservative force. Work done by an external force (equal and opposite to the electrostatic force) in bringing a charge q from a point R to a point P is $V_P - V_R$, which is the difference in potential energy of charge q between the final and initial points.
2. Potential at a point is the work done per unit charge (by an external agency) in bringing a charge from infinity to that point. Potential at a point is arbitrary to within an additive constant, since it is the potential difference between two points which is physically significant. If potential at infinity is chosen to be zero; potential at a point with position vector **r** due to a point charge Q placed at the origin is given is given by
$$\boldsymbol{V(r) = \frac{Q}{4\pi\varepsilon_0 r}}$$
3. The electrostatic potential at a point with position vector **r** due to a point dipole of dipole moment **p** placed at the origin is
$$\boldsymbol{V(r) = \frac{1}{4\pi\varepsilon_0}\frac{p.\hat{r}}{r^2}}$$
The result is true also for a dipole (with charges $-q$ and q separated by $2a$) for $r >> a$.
4. For a charge configuration q_1, q_2, ..., q_n with position vectors $\mathbf{r}_1$, $\mathbf{r}_2$, ... $\mathbf{r}_n$, the potential at a point P is given by the superposition principle
$$\boldsymbol{V = \frac{1}{4\pi\varepsilon_0}\left(\frac{q_1}{r_{1P}} + \frac{q_2}{r_{2P}} + \cdots + \frac{q_n}{r_{nP}}\right)}$$
where $\boldsymbol{r_{1P}}$ is the distance between $\boldsymbol{q_1}$ and P, as and so on.
5. An equipotential surface is a surface over which potential has a constant value. For a point charge, concentric spheres centered at a location of the charge are equipotential surfaces. The electric field **E** at a point is perpendicular to the equipotential surface through the point. **E** is in the direction of the steepest decrease of potential.
6. Potential energy stored in a system of charges is the work done (by an external agency) in assembling the charges at their locations. Potential energy of two charges $\boldsymbol{q_1}$, $\boldsymbol{q_2}$ at $\boldsymbol{r_1}$, $\boldsymbol{r_2}$ is given by
$$\boldsymbol{U = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r_{12}}}$$
where $\boldsymbol{r_{12}}$ is distance between $\boldsymbol{q_1}$ and $\boldsymbol{q_2}$
7. The potential energy of a charge q in an external potential $V(\mathbf{r})$ is $qV(\mathbf{r})$. The potential energy of a dipole moment **p** in a uniform electric field **E** is $-\mathbf{p}.\mathbf{E}$.
8. Electrostatics field **E** is zero in the interior of a conductor; just outside the surface of a charged conductor, **E** is normal to the surface given by $\boldsymbol{E = \frac{\sigma}{\varepsilon_0}\hat{n}}$, where $\hat{\boldsymbol{n}}$ is the unit vector along the outward normal to the surface and σ is the surface charge density. Charges in a conductor can reside only at its surface. Potential is constant within and on the surface of a conductor. In a cavity within a conductor (with no charges), the electric field is zero.
9. A capacitor is a system of two conductors separated by an insulator. Its capacitance is defined by $C = Q/V$, where Q and $-Q$ are the charges on the two conductors and V is the potential difference between them. C is determined purely geometrically, by the shapes, sizes and relative positions of the two conductors.
The unit of capacitance is farad: 1 F = 1 C V^{-1}.

For a parallel plate capacitor (with vacuum between the plates),

$$C = \varepsilon_0 \frac{A}{d}$$

where A is the area of each plate and d the separation between them.

10. If the medium between the plates of a capacitor is filled with an insulating substance (dielectric), the electric field due to the charged plates induces a net dipole moment in the dielectric. This effect, called polarisation, gives rise to a field in the opposite direction. The net electric field inside the dielectric and hence the potential difference between the plates is thus reduced. Consequently, the capacitance C increases from its value C_0 when there is no medium (vacuum),

    $C = KC_0$

    where K is the dielectric constant of the insulating substance.
11. For capacitors in the series combination, the total capacitance C is given by

    $$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \cdots$$

    In the parallel combination, the total capacitance C is:

    $$C = C_1 + C_2 + C_3 + \ldots$$

    where C_1, C_2, C_3... are individual capacitances.
12. The energy U stored in a capacitor of capacitance C, with charge Q and voltage V is

    $$U = \frac{1}{2}QV = \frac{1}{2}CV^2 = \frac{1}{2}\left(Q^2/C\right)$$

    The electric energy density (energy per unit volume) in a region with electric field is $(1/2)\varepsilon_0 E^2$.
13. A Van de Graaff generator consists of a large spherical conducting shell (a few metre in diameter). By means of a moving belt and suitable brushes, charge is continuously transferred to the shell and potential difference of the order of several million volts is built up, which can be used for accelerating charged particles.

Physical quantity	**Symbol**	**Dimensions**	**Unit**	**Remark**
Potential	*Φ or V*	$[M^1L^2T^{-3}A^{-1}]$	V	Potential difference is physically significant
Capacitance	C	$[M^{-1}L^{-2}T^{-4}A^2]$	F	
Polarisation	P	$[L^{-2}AT]$	Cm^{-2}	Dipole moment per unit volume
Dielectric constant	K	Dimensionless		

VERY SHORT ANSWER QUESTIONS (2 MARKS)

1. Can there be electric potential at a point with zero electric intensity ? Give an example.
2. Can there be electric intensity at a point with zero electric potential ? Give an example.
3. What are meant by equipotential surfaces?
4. Why is the electric field always at right angles to the equipotential surface? Explain.
5. Three capacitors of capacitances 1 F, 2 μF, and 3 μF are connected in parallel.
   (a) What is the ratio of charges? (b) What is the ratio of potential differences?
6. Three capacitors of capacitances 1 F, 2 μF, and 3 μF are connected in series.
   (a) What is the ratio of charges? (b) What is the ratio of potential differences?
7. What happens to the capacitance of a parallel plate at capacitor if the area of its plates is doubled?

SHORT ANSWER QUESTIONS (4 MARKS)

1. Derive an expression for the electric potential due to a point charge.
2. Derive an expression for the potential energy of an electric dipole placed in a uniform electric field.
3. Derive an expression for the capacitance of a parallel plate capacitor.
4. Explain the behaviour of dielectrics in an external field.

LONG ANSWER QUESTIONS (8 MARKS)

1. Define electric potential. Derive an expression for the electric potential due to an electric dipole and hence the electric potential at a point (a) the axial line of electric dipole (b0 on the equatorial line of electric dipole.
2. Explain series and parallel combination of capacitors. Derive the formula for equivalent capacitance in each combination.
3. Derive an expression for the energy stored in a capacitor. What is the energy stored when the space between the plates is filled with a dielectric
   (a) With charging battery disconnected?
   (b) With charging battery connected in the circuit ?

Chapter 6

CURRENT ELECTRICITY

6.1 INTRODUCTION

In Chapter 4, all charges whether free or bound, were considered to be at rest. Charges in motion constitute an electric current. Such currents occur naturally in many situations. Lightning is one such phenomenon in which charges flow from the clouds to the earth through the atmosphere, sometimes with disastrous results. The flow of charges in lightning is not steady, but in our everyday life we see many devices where charges flow in a steady manner, like water flowing smoothly in a river. A torch and a cell-driven clock are examples of such devices. In the present chapter, we shall study some of the basic laws concerning steady electric currents.

6.2 ELECTRIC CURRENT

Image a small area held normal to the direction of flow of charge. Both the positive and the negative charges may flow forward and backward across the area. In a given time interval t, let q_+, be the net amount (i.e., forward *minus* backward) of positive charge that flows in the forward direction across the area. Similarly, let q_- be the net amount of negative charge flowing across the area in the forward direction. The net amount of charge flowing across the area in the forward direction in the time interval t, then, is $q = q_+ - q_-$ This is proportional to t for steady current and the quotient

$$I = \frac{q}{t} \tag{6.1}$$

is defined to be the *current* across the area in the forward direction. (If it turn out to be a negative number, it implies a current in the backward direction).

Currents are not always steady and hence more generally. We define the current as follows. Let ΔQ be the net charge flowing across a cross-section of a conductor during the time interval Δt [i.e., between times t and (t + Δt)]. Then, the current at time t across the cross-section of the conductor is defined as the value of the ratio of ΔQ to Δt in the limit of Δt tending to zero.

$$I(t) \equiv \lim_{\Delta t \to 0} \frac{\Delta Q}{\Delta t} \tag{6.2}$$

In SI units, the unit of current is ampere. An ampere is defined through magnetic effects of currents that we will study in the following chapter. An ampere is typically the order of magnitude of currents in domestic appliances. An average lightning carries currents of the order of tens of thousands of amperes and at the other extreme, currents in our nervers are in microamperes.

6.3 ELECTRIC CURRENTS IN CONDUCTORS

An electric charge will experience a force if an electric field is applied. If it is free to move, it will thus move contributing to a current. In nature, free charged particles do exist like in upper strata of atmosphere called the *ionosphere*. However, in atom and molecules, the negatively charged electrons and the positively charged nuclei are bound to each other and are thus not free to move. Bulk matter is made up of many molecules, a gram of water, for example, contains approximately 10^{22} molecules. These molecules are so closely packed that the electrons are no longer attached to individual nuclei. In some materials, the electrons will still be bound, i.e., they will not accelerate even if an electric field is applied. In other materials, notably metals, some of the electrons are practically free to move within the bulk

material. These materials, generally called conductors, develop electric currents in them when an electric field is applied.

If we consider solid conductors, then of course the atoms are tightly bound to each other so that the current is carried by the negatively charged electrons. There are, however, other types of conductors like electrolytic solutions where positive and negative charges both can move. In our discussions, we will focus only on solid conductors so that the current is carried by the negatively charged electrons in the background of fixed positive ions.

Consider first the case when no electric field is present. The electrons will be moving due to thermal motion during which they collide with the fixed ions. An electron colliding with an ion emerges with the same speed as before the collision. However, the direction of its velocity after the collision is completely random. At a given time, there is no preferential direction for the velocities of the electrons. Thus on the average the number of electrons travelling in any direction will be equal to the number of electrons travelling in the opposite direction. So, there will be no net electric current.

Fig. 6.1 Charges +Q and –Q put at the ends of a metallic cylinder. The electrons will drift because of the electric field created to neutralise the charges. The current thus will stop after a while unless the charges +Q and –Q are continuously replenished.

Let us now see what happens to such a piece of conductor if an electric field is applied. To focus our thoughts, imagine the conductor in the shape of a cylinder of radius R (Fig. 6.1). Suppose we now take two thin circular discs of a dielectric of the same radius and put positive charge +Q distributed over one disc and similarly –Q at the other disc. We attach the two discs on the two flat surfaces of the cylinder. An electric field will be created and is directed from the positive towards the negative charge. The electrons will be accelerated due to this field towards +Q. They will thus move to neutralise the charges. The electrons, as long as they moving, will constitute an electric current. Hence in the situation considered, there will be a current for a very short while and no current thereafter.

We can also imagine a mechanism where the ends of the cylinder are supplied with fresh charges to make up for any charges neutralised by electrons moving inside the conductor. In that case, there will be a steady electric field in the body of the conductor. This will result in a continuous current rather than a current for a short period of time. Mechanisms, which maintain a steady electric field are cells or batteries that we shall study later in this chapter. In the next sections, we shall study the steady current that results from a steady electric field in conductors.

6.4 OHM'S LAW

A basic law regarding flow of currents was discovered by G. S. Ohm in 1828, long before the physical mechanism responsible for flow of currents was discovered. Imagine a conductor through which a current I is flowing and let V be the potential difference between the ends of the conductor. Then Ohm's law states that

$$V \propto I$$

Or,
$$V = R\,I \qquad (6.3)$$

where the constant of proportionally R is called the *resistance* of the conductor. The SI unit of resistance are ohm, and is denoted by the symbol Ω. The resistance R not only depends on the material of the conductor but also on the dimensions of the conductor. The dependence of R on the dimensions of the conductor can easily be determined as follows.

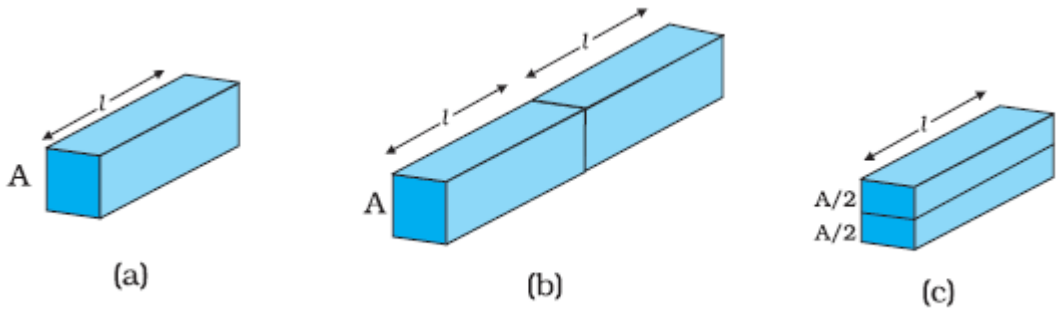


Fig. 6.2 Illustrating the relation R = ρ*l*/*A* for a rectangular slab of length *l* and area of cross-section *A*.

Consider a conductor satisfying Eq.(6.3) to be in the form of a slab of length l and cross sectional area A [Fig.6.2(a)]. Imagine placing two such identical slabs side by side [Fig. 6.2 (b)], so that the length of the combination is 2*l*. The current flowing through the combination is the same as that flowing through other of the slabs. If V is the potential difference across the ends of the first slab, then V is also the potential difference across the ends of the second slab since the second slab is identical to the first and the same current I flows through both. The potential difference across the ends of the combination is clearly sum of the [potential difference across the two individual slabs and hence equals 2 V. The current through the combination is I and the resistance of the combination R_C is [from Eq. (6.3)].

$$R_C = \frac{2V}{I} = 2R \qquad (6.4)$$

Since V/I = R, the resistance of either of the slabs. Thus, doubling the length of a conductor doubles the resistance. In general, then resistance is proportional to length,

$$R \propto l \qquad (6.5)$$

Next, imagine dividing the slab into two by cutting it lengthwise so that the slab can be considered as a combination of two identical slabs of length l , but each having a cross sectional area of A/2 [Fig. 6.2(c)].

For a given voltage V across the slab, if I is the current through the entire slab, then clearly the current flowing through each of the two half-slabs is I/2. Since the potential difference across the ends of the half-slabs is V, i.e., the same as across the full slab, the resistance of each of the half-slab R_1 is

$$R_1 = \frac{V}{(I/2)} = 2\frac{V}{I} = 2R \qquad (6.6)$$

Thus, halving the area of the cross-section of a conductor doubles the resistance. In general, then the resistance R is inversely proportional to the cross-sectional area.

$$R \propto \frac{1}{A} \qquad (6.7)$$

Combining Eqs. (6.5) and (6.7), we have

$$R \propto \frac{l}{A} \qquad (6.8)$$

and hence for a given conductor, $R \propto \rho \frac{l}{A}$ (6.9)

where the constant of proportionality ρ depends on the material of the conductor but not its dimensions. ρ is called *resistivity*.

Using the last equation, Ohm's law reads

$$V = I \times R = \frac{I \rho l}{A} \quad (6.10)$$

Georg Simon Ohm (1787–1854) German physicist, professor at Munich. Ohm was led to his law by an analogy between the conduction of heat: the electric field is analogous to the temperature gradient, and the electric current is analogous to the heat flow.

Current per unit area (taken normal to the current), I/A, is called *current density* and is denoted by j. The SI units of the current density are A/m^2. Further , if E is the magnitude of uniform electric field in the conductor whose length is l, then the potential difference V across its ends is El. Using these, the last equation reads.

$$E\, l = j\, \rho\, l$$

Or , $$E = j\, \rho \quad (6.11)$$

The above relation for magnitudes ***E*** and ***J*** can indeed be cast in a ***vector*** form. The current density, (which we have defined as the current through unit area ***normal*** to the current) ia also directed along ***E,*** and is also a vector $\boldsymbol{j}\ (\equiv \boldsymbol{j}\,\boldsymbol{E}/E)$. Thus, the last equation can be written as,

$$\boldsymbol{E} = \boldsymbol{j}\, \rho \quad (6.12)$$

Or, $$\boldsymbol{j} = \sigma\, \boldsymbol{E} \quad (6.13)$$

where $\sigma \equiv 1/\rho$ is called the *conductivity*. Ohm's law is often stated in an equivalent form, Eq.(6.13) in addition to Eq.(6.3) . In the next section, we will try to understand the origin of the Ohm's law as arising from the characteristics of the drift of electrons.

6.5 DRIFT OF ELECTRONS AND THE ORIGIN OF RESISTIVITY

As remarked before, an electron will suffer collisions with the heavy fixed ions, but after collision, it will emerge with the same speed but in random directions. If we consider all the electrons, their average velocity will be zero since their directions are random. Thus, if there are ***N*** electrons and the velocity of the $\boldsymbol{i}^{th}$ electron ($\boldsymbol{i}$ = 1, 2, 3, ...N) at a given time is $\mathbf{v}_1$, then

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{v_t} = \mathbf{0} \quad (6.14)$$

Consider now the situation when an electric field is present. Electrons will be accelerated due to this field by

$$\boldsymbol{a} = \frac{-e\,\boldsymbol{E}}{\boldsymbol{m}} \quad (6.15)$$

where **-e** is the charge and **m** is the mass of an electron. Consider again the $\mathbf{i}^{th}$ electron at a given time **t**. This electron would have had its last collision some time before **t**, and let $\mathbf{t_i}$ be

the time elapsed after its last collision. If v_i was its velocity immediately after the last collision, then its velocity V_i at time t is

$$V_i = v_i + \frac{-eE}{m} t_i \qquad (6.16)$$

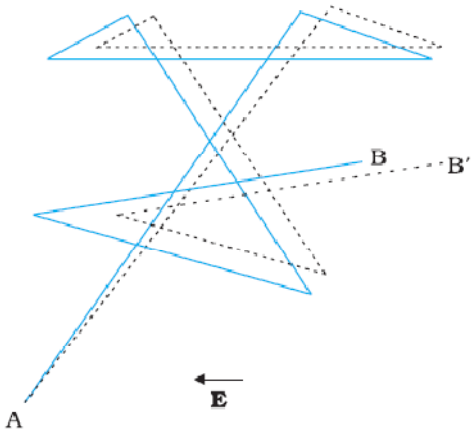


Fig. 6.3 A schematic picture of an electron moving from a point A to another point B through repeated collisions, and straight line travel between collisions (full lines). If an electric field is applied as shown, the electron ends up at point B′ (dotted lines). A slight drift in a direction opposite the electric field is visible.

Since starting with its last collision it was accelerated (Fig. 6.3) with an acceleration given by Eq. (6.15) for a time interval $\mathbf{t_i}$. The average velocity of the electrons at time t is the average of all the V_i**'s.** The average of v_i**'**s is zero [Eq.(6.14)] since immediately after any collision, the direction of the velocity of an electron is completely random. The collisions of the electrons do not occur at regular intervals but at random times. Let us denote by, the average time between successive collisions. Then at a given time, some of the electrons would have spent time more than τ and some less than . In other words, the time $\mathbf{t_i}$ in Eq. (6.16) will be less than τ for some and more than τ for others as we go through the values of i = 1, 2 ,..... N. The average value of $\mathbf{t_i}$ then is τ (known as *relaxation time*).

Thus, averaging Eq. (6.16) over the N-electrons at any given time t gives us for the average velocity $\mathbf{v_d}$

$$v_d = (V_i)_{average} = (v_i)_{average} - \frac{eE}{m} (t_i)_{average}$$
$$= 0 - \frac{eE}{m}\tau = -\frac{eE}{m}\tau \qquad (6.17)$$

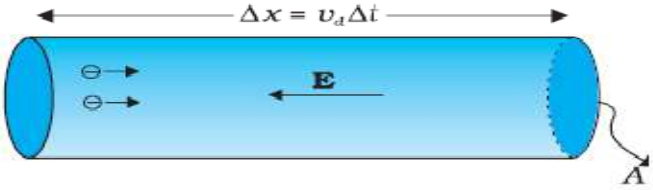


Fig. 6.4 Current in a metallic conductor. The magnitude of current density in a metal is the magnitude of charge contained in a cylinder of unit area and length v_d.

This last result is surprising. It tells us that the electrons move with an average velocity which is independent of time, although electrons are accelerated. This is the phenomenon of drift and the velocity v_d in Eq.(6.17) is called the ***drift velocity***.

Because of the draft, there will be net transport of charges across any area perpendicular to E. Consider a planer area A, located inside the conductor such that the normal to the area is parallel to **E** (Fig. 6.4). Then because of the drift, in an infinitesimal amount of time Δt, all electrons to the left of the area at distances upto $|v_d|\Delta t$ would have crossed the area. If n is the number of free electrons per unit volume in the metal, then there are $\mathbf{n}\ \Delta t\ |V_d|A$ such electrons. Since each electron carries a charge **– e**, the total charge transported across this area **A** to the right in time Δt $is - ne\ A|v_d|\Delta t$. **E** is directed towards the left and hence the total charge transported along **E** across the area is negative of this. The amount of charge crossing the area **A** in the Δt is by definition [Eq.(6.2)] $\mathbf{I}\ \Delta t$ **,** where **I** is the magnitude of the current. Hence,

$$I\ \Delta t + ne\ A|v_d|\Delta t. \tag{6.18}$$

Substituting the value of $|v_d|$from Eq. (6.17)

$$I\Delta t = \frac{e^2 A}{m}\tau n\ \Delta t\ |E| \tag{6.19}$$

By definition **I** is related to the magnitude $|j|$ of the current density by

$$\mathbf{I} = |j|\ A \tag{6.20}$$

Hence, from Eqs. (6.19) and (6.20),

$$|j| = \frac{ne^2}{m}\ \tau\ |E| \tag{6 .21}$$

The vector **j** is parallel to **E** and hence we can write Eq . (6. 21) in the vector from

$$j = \frac{ne^2}{m}\ \tau\ E \tag{6.22}$$

Comparison with Eq. (3.13) shows that Eq. (6.22) is exactly the Ohm's law, if we identify the conductivity σ as

$$\sigma = \frac{ne^2}{m}\tau \tag{6.23}$$

We thus see that a very simple picture of electrical conduction reproduces Ohm's law. We have, of course, made assumptions that τ and n are constants, independent of E. We shall, in the next section, discuss the limitations of Ohm's law.

6.5.1 Mobility

As we have seen, conductivity arises from mobile charge carriers. In metals, these mobile charge carries are electrons; in an ionised gas, they are electrons and positive charged ions; in an electrolyte, these can be both positive and negative ions.

An important quantity is the *mobility* μ defined as the magnitude of the draft velocity per unit electric field:

$$\mu = \frac{|v_d|}{E} \tag{6.24}$$

The SI unit of mobility is m^2 / Vs and is 10^4 of the mobility in practical units (Cm^2 / VS) . Mobility is positive. From Eq. (6.17), we have

$$v_d = \frac{e\tau E}{m}$$

Hence,

$$\mu = \frac{v_d}{E} = \frac{e\tau}{m} \tag{6.25}$$

where τ is the average collision time for electrons.

6.6 LIMITATIONS OF OHM'S LAW

Although Ohm's law has been found valid over a large class of materials, there do exist materials and devices used in electric circuits where the proportionality of *V* and *I* does not hold. The deviations broadly are one or more of the following types:

(a) V ceases to be proportional to 1 (Fig . 6.5).

(b) The relation between V and I depends on the sign of V. In other wards, if I is the current for a certain V, then reversing the direction of V keeping its magnitude fixed, does not produce a current of the same magnitude as I in the opposite direction (Fig. 6.6). This happens, for examples, in a diode which we will study in Chapter 15.

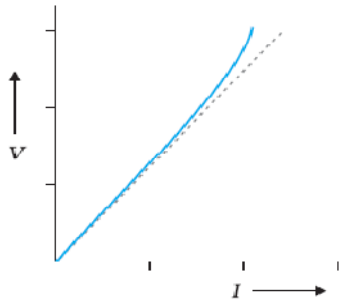


Fig. 6.5 The dashed line represents the linear Ohm's law. The solid line is the voltage *V* versus current *I* for a good conductor.

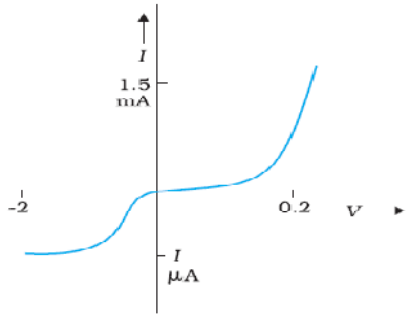


Fig. 6.6 Characteristic curve of a diode. Note the different scales for negative and positive values of the voltage and current.

(c) The relation between V and I is not unique, i.e., there is more than one value of V for the same current I (Fig . 6.7). A material exhibiting such behaviour is Gas.

Material and devices not obeying Ohm's Law in the form of Eq.(6.3) are actually widely used in electronic circuits. In this and a few subsequent chapters, however, we will study the electrical currents in materials that obey Ohm's law.

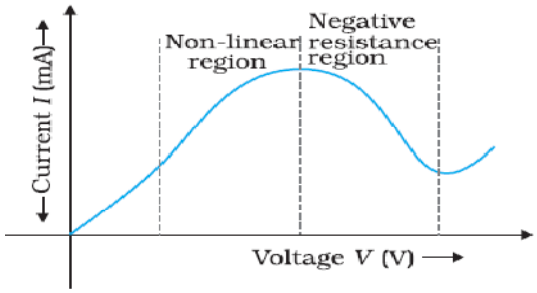


Fig. 6.7 Variation of current versus voltage for GaAs.

6.7 RESISTIVITY OF VARIOUS MATERIALS

The resistivities of various common materials are listed in Table 6.1. The materials are classified as conductors, semiconductors and insulators depending on their resistivities, in an increasing order of their values. Metals have low resistivities in the range of $10^{-8}\ \Omega m$ to $10^{-6}\ \Omega m$. At the other end are insulators like ceramic, rubber and plastics having resistivities 10^{18} times greater than metals or more. In between the two characteristically decreasing with a rise in temperature. The resistivities of semiconductors are also affected by presence of small amount of impurities. This last feature is exploited in use of semiconductors for electronic devices.

Table 6.1 Resistivities of some materials

Material	**Resistivity,** $\rho\ (\Omega\, m)$ at $0°C$	**Temperature coefficient of resistivity,** $\alpha\ (°C)^{-1}$ $\frac{1}{\rho}\left(\frac{d\rho}{dT}\right)$ at $0°C$
Conductors		
Silver	1.6×10^{-8}	0.0041
Copper	1.7×10^{-8}	0.0068
Aluminium	2.7×10^{-8}	0.0043
Tungsten	5.6×10^{-8}	0.0045
Iron	10×10^{-8}	0.0065
Platinum	11×10^{-8}	0.0039
Mercury	98×10^{-8}	0.0009
Nichrome	$\sim 100 \times 10^{-8}$	0.0004
(alloy of Ni, Fe, Cr) Manganin (alloy)	48×10^{-8}	0.002×10^{-3}
Semiconductors		
Carbon (graphite)	3.5×10^{-5}	– 0.0005
Germanium	0.46	– 0.05
Silicon	2300	– 0.07
Insulators		
Pure Water	2.5×10^{5}	
Glass	$10^{10} – 10^{14}$	
Hard Rubber	$10^{13} – 10^{16}$	
NaCl	$\sim 10^{14}$	
Fused Quartz	$\sim 10^{16}$	

Commercially produced resistors for domestic use or in laboratories are of two major types : *wire bound resistors and carbon resistors.* Wire bound resistors are made by winding the wires of an alloy, viz., manganin, constantan, nichrome or similar ones. The choice of these materials is dictated mostly by the fact that their resistivities are relatively insensitive to temperature. These resistances are typically in the range of a fraction of an ohm to a few hundred ohms.

Resistors in the higher range are made mostly from carbon. Carbon resistors are compact, inexpensive and thus find extensive use in electronic circuits. Carbon resistors are small in size and hence their values are given using a colour code.

Table 6.2 Resistor Colour Codes

Color	Number	Multiplier	Tolerance (%)
Black	0	1	
Brown	1	10^1	
Red	2	10^2	
Orange	3	10^3	
Yellow	4	10^4	
Green	5	10^5	
Blue	6	10^6	
Violet	7	10^7	
Gray	8	10^8	
White	9	10^9	
Gold		10^{-1}	5
Silver		10^{-2}	10
No colour			20

The resistors have a set of co-axial coloured rings on them whose significance are listed in Table 6.2. The first two bands from the end indicate the first two significant figures of the resistance in ohms. The third band indicates the decimal multiplier (as listed in Table 6.2).

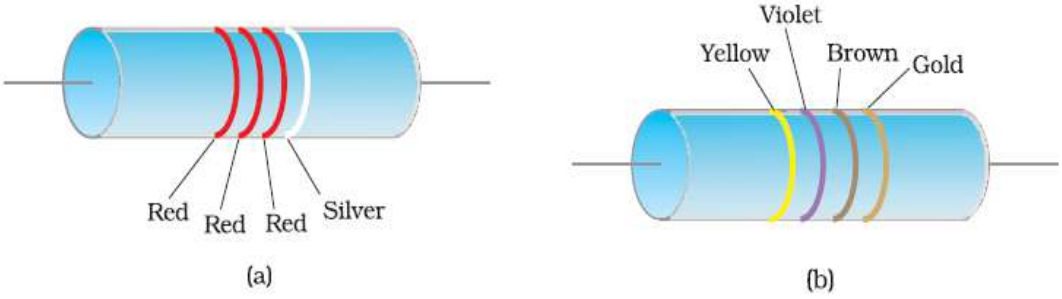


Fig. 6.8 Colour coded resistors (a) $(22 \times 10^2\ \Omega) \pm 10\%$, (b) $(47 \times 10\ \Omega) \pm 5\%$.

The last band stands for tolerance or possible variation in percentage about the indicated values. Sometimes, this last band is absent and that indicates a tolerance of 20% (Fig. 6.8). For example, if the four colours are orange, blue, yellow and gold, the resistance value is$36 \times 10^4\ \Omega$, with a tolerance value of 5%.

6.8 TEMPERATURE DEPENDENCE OF RESISTIVITY

The resistivity of a material is found to be dependent on the temperature. Different materials do not exhibit the same dependence on temperatures. Over a limited range of temperatures, that is not too large, the resistivity of a metallic conductor is approximately given by,

$$\boldsymbol{\rho_r = \rho_0[1 + \alpha(T - T_0)]} \qquad (6.26)$$

where $\boldsymbol{\rho_r}$ is the resistivity at a temperature T and $\boldsymbol{\rho_0}$ is the same at a reference temperature $\boldsymbol{T_0}$. α is called the temperature co-efficient of resistivity, and from Eq. (6.26), the dimension of α is (Temperature)$^{-1}$.

For metals, $\boldsymbol{\alpha}$ is positive and values of $\boldsymbol{\alpha}$ for some metals at $\boldsymbol{T_0}$ = 0^0C are listed in Table 6.1.

The relation of Eq.(6.26) implies that a graph of $\boldsymbol{\rho_r}$ plotted against T would be a straight line.

At temperatures much lower than 0^0C, the graph, however, deviates considerably from a straight line (Fig. 6.9).

Equation (6.26) thus, can be used approximately over a limited range of T around any reference temperature$\boldsymbol{T_0}$, where the graph can be approximated as a straight line.

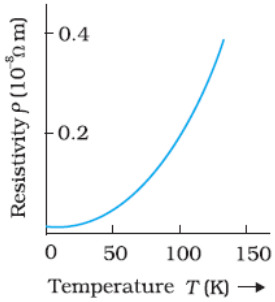


Fig. 6.9 Resistivity $\boldsymbol{\rho_r}$ of copper as a function of temperature T.

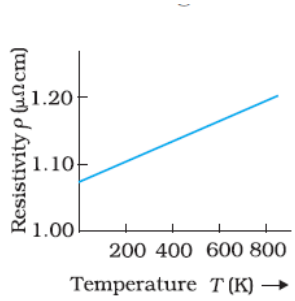


Fig. 6.10 Resistivity $\boldsymbol{\rho_r}$ of nichrome as a function of absolute temperature T.

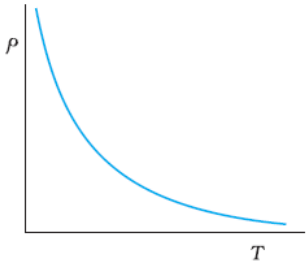


Fig. 6.11 Temperature dependence of resistivity for a typical emiconductor.

Some materials like Nichrome (which is an alloy of nickel, iron and chromium) exhibit a very weak dependence of resistivity with temperature (Fig . 6.10). Manganin and constantan have similar properties. These materials are thus widely used in wire bound standard resistors since their resistance values would change very little with temperatures.

Unlike metals, the resistivity of semiconductors decrease with increasing temperatures, A typical dependence is shown in Fig 6.11.

We can qualitatively understand the temperature dependence of resistivity, in the light of our derivation of Eq.(6.23). From this equation, resistivity of a material is given by

$$\boldsymbol{\rho = \frac{1}{\sigma} = \frac{m}{ne^2\tau}} \qquad (6.27)$$

$\boldsymbol{\rho}$ thus depends inversely both on the number n of free electrons per unit volume and on the average time $\boldsymbol{\tau}$ between collisions. As we increase temperature, average speed of the electrons, which act as the carriers of current, increases resulting in more frequent collisions. The average time of collisions**,** thus decreases with temperature.

In a metal, n is not dependent on temperature to any appreciable extent and thus the decrease in the value of $\boldsymbol{\tau}$ with rise in temperature causes $\boldsymbol{\rho}$ to increase as we have observed.

For insulators and semiconductors, however, n increases with temperature. This increase more than compensates any decrease in $\boldsymbol{\tau}$ in Eq.(6.23) so that for such materials, $\boldsymbol{\rho}$ decreases with temperature.

6.9 ELECTRICAL ENERGY, POWER

Consider a conductor with end points A and B, in which a current I is flowing from A to B. The electric potential at A and B are denoted by V(A) and V(B) respectively. Since current is flowing from A to B, V(A) > V(B) and the potential different across AB is V = V(A) – V(B) > 0.

In a time interval Δt , an amount of charge $\Delta Q = I\,\Delta t$ travels from A to B. The potential energy of the charge at A, by definition , was Q V(A) and similarly at B, it is Q V(B). Thus, change in its potential energy ΔU_{pot} is

ΔU_{pot} = Final potential energy – Initial potential energy

$$= \Delta Q\,[(v(B) - V(A)] = \Delta Q\,V$$

$$= -\,I\,V\Delta t < 0 \qquad (6.28)$$

If charges moved without collisions through the conductor, their kinetic energy would also change so that the total energy is uncharged. Conservation of total energy would then imply that.

$$\Delta K = -\,\Delta U_{\text{pot}} \qquad (6.29)$$

That is,

$$\Delta K = I\,V\Delta t > 0 \qquad (6.30)$$

Thus, in case charges were moving freely through the conductor under the action of electric field, their kinetic energy would increase as they move. We have, however, seen earlier that on the average, charge carriers do not move with acceleration but with a steady drift velocity. This is because of the collisions with ions and atoms during transit. During collisions, the energy gained by the charges thus is shared with the atoms. The atoms vibrate more vigorously, i.e., the conductor heats up. Thus, in an actual conductor, an amount of energy dissipated as heat in the conductor during the time interval Δt is,

$$\Delta W = I\,V\Delta t \qquad (6.31)$$

The energy dissipated per unit time is the power dissipated P = $\Delta W / \Delta t$ and we have,

$$P = I\,V \qquad (6.32)$$

Using Ohm's law V = IR , we get

$$P = I^2\,R = V^2 / R \qquad (6.33)$$

As the power loss ("ohmic loss") in a conductor of resistance R carrying a current I. It is this power which heats up, for example, the coil of an electric bulb to incandescence, radiating out heat and light.

Where does the power come from? As we have reasoned before, we need an external source to keep a steady current through the conductor. It is clearly this source which must supply this power. In the simple circuit shown with a cell (Fig. 6.12), it is the chemical energy of the cell which supplies this power for as long as it can.

The expressions for power, Eqs, (6.32) and (6.33), show the dependence of the power dissipated in a resistor R on the current through it and the voltage across it. Equation (6.33) has an important application to power transmission. Electrical power is transmitted from power stations to homes factories, which may be hundreds of miles away, via transmission cables.

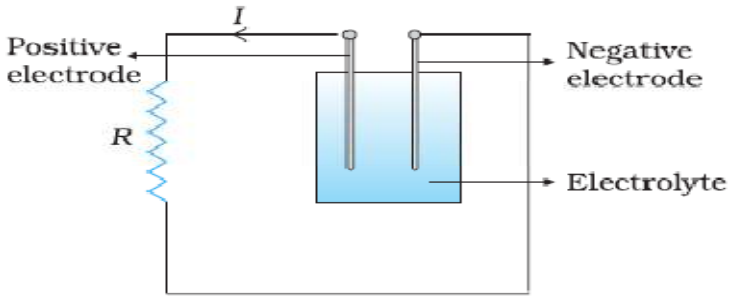


Fig. 6.12 Heat is produced in the resistor *R* which is connected across the terminals of a cell. The energy dissipated in the resistor *R* comes from the chemical energy of the electrolyte.

One obviously wants to minimise the power loss in the transmission cables connecting the power stations to homes and factories. We shall see now how this can be achieved. Consider a device R, to which a power P is to be delivered via transmission cables having a resistance R_C to be dissipated by it finally. If V is the voltage across R and I the current through it, then

$$P = V I \tag{6.34}$$

The connecting wires from the power station to the device has a finite resistance R_C . The power dissipated in the connecting wires, which is wasted is P_C with

$$P_C = I^2 R_C = \frac{P^2 R_C}{V^2} \tag{6.35}$$

From Eq.(6.32). Thus, to drive a device of power P, the power wasted in the connecting wires in inversely proportional to V^2. The transmission cables from power stations are hundreds on miles long and their resistance R_C is considerable. To reduce P_C, these wires carry current at enormous values of V and this is the reason for the high voltage danger signs on transmission lines – a common sight as we move away from populated areas. Using electricity at such voltages is not safe and hence at the other end, a device called a transformer lowers the voltage to a value suitable for use.

6.10 COMBINATION OF RESISTORS – SERIES AND PARALLEL

The current through a single resistor R across which there is a potential difference V is given by Ohm's law I = V/R. Resistors are sometimes joined together and there are simple rules for calculation of equivalent resistance of such combination.

Fig. 6.13 A series combination of two resistor R_1 and R_2.

Two resistors are said to be in *series* if only one of their end points is joined (Fig. 6.13). If a third resistor is joined with the series combination of the two (Fig.6.14), then all three are said to be in series. Clearly, we can extend this definition to series combination of any number of resistors.

Fig. 6.14 A series combination of three resistors R_1, R_2, R_3.

Two or more resistors are said to be in *parallel* if one end of all the resistors is joined together and similarly the other ends joined together (Fig.6.15).

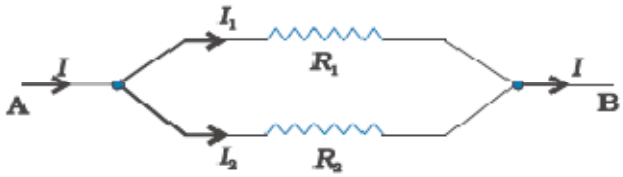


Fig. 6.15 Two resistors R_1 and R_2 connected in parallel.

Consider two resistors R_1 and R_2 in series. The charge which leaves R_1 must be entering R_2. Since current measures the rate of flow of charge, this means that the same current I flows through R_1 and R_2. By Ohm's law:

Potential difference across $R_1 = V_1 = IR_1$, and

Potential difference across $R_2 = V_2 = IR_2$.

The potential difference V across the combination is $V_1 + V_2$. Hence,

$$V = V_1 + V_2 = I (R_1 + R_2) \quad (6.36)$$

This is as if the combination had an equivalent resistance R_{eq} , which by Ohm's law is

$$R_{eq} = \frac{V}{I} = (R_1 + R_2) \quad (6.37)$$

If we had three resistors connected in series, then similarly

$$V = I R_1 + I R_2 + I R_3 = I (R_1 + R_2 + R_3) \quad (6.38)$$

This obviously can be extended to a series combination of any number n of resistors R_1 ,R_2 ....., R_n. The equivalent resistance R_{eq} is

$$R_{eq} = R_1 + R_2 + \ldots\ldots + R_n \quad (6.39)$$

Consider now the parallel combination of two resistors (Fig.6.15). The charge that flows in at A from the left flows out partly through R_1 and partly through R_2. The current I, I_1, I_2 shown in the figure are the rates of flow of charge at the points indicated. Hence,

$$I = I_1 + I_2 \quad (6.40)$$

The potential difference between A and B is given by the Ohm's law applied to R_1

$$V = I_1 R_1 \quad (6.41)$$

Also, Ohm's law applied to R_2 gives

$$V = I_2 R_2 \quad (6.42)$$

$$\therefore I = I_1 + I_2 = \frac{V}{R_1} + \frac{V}{R_2} = V\left(\frac{1}{R_1} + \frac{1}{R_2}\right) \quad (6.43)$$

If the combination was replaced by an equivalent resistance R_{eq} , we would have, by Ohm's law

$$I = \frac{V}{R_{eq}} \quad (6.44)$$

Hence,

$$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} \quad (6.45)$$

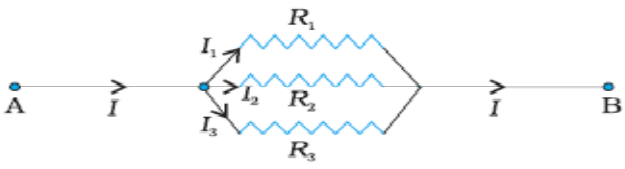


Fig. 6.16 Parallel combination of three resistors R_1, R_2 and R_3.

We can easily see how this extends to three resistors in parallel (Fig.6.16)

Exactly as before

$$I = I_1 + I_2 + I_3 \tag{6.46}$$

And applying Ohm's law to R_1, R_2 and R_3 we get,

$$V = I_1 R_1 \ , \ V = I_2 R_2 \ , \ V = I_3 R_3 \tag{6.47}$$

So that

$$I = I_1 + I_2 + I_3 = V\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right) \tag{6.48}$$

An equivalent resistance R_{eq} that replaces the combination, would be such that

$$I = \frac{V}{R_{eq}} \tag{6.49}$$

And hence

$$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \tag{6.50}$$

We can reason similarly for any number of resistors in parallel. The equivalent resistance of n resistors R_1, R_2 ....., R_n is

$$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} + \cdots \ldots + \frac{1}{R_n} \tag{6.51}$$

These formulae for equivalent resistance can be used to find out currents and voltage in more complicated circuits.

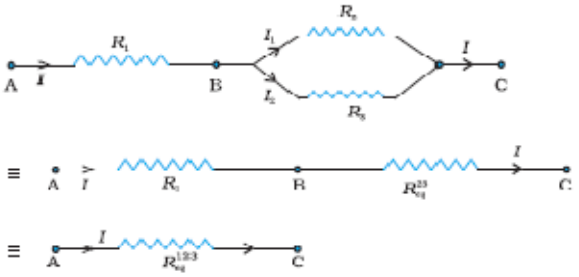


Fig. 6.17 A combination of three resistors R_1, R_2 and R_3. R_2, R_3 are in parallel with an equivalent resistance R_{eq}^{23}. R_1 and R_{eq}^{23} are in series with an equivalent resistance R_{eq}^{123}

Consider for example, the circuit in Fig. (6.17), where there are three resistors R_1, R_2 and R_3. R_2 and R_3 are in parallel and hence we can replace them by an equivalent R_{eq}^{23} between point B and C with

$$\frac{1}{R_{eq}^{23}} = \frac{1}{R_2} + \frac{1}{R_3}$$

Or,
$$R_{eq}^{23} = \frac{R_2 R_3}{R_2 + R_3} \tag{6.52}$$

The circuit now has R_1 and R_{eq}^{23} in series and hence their combination can be replaced by an equivalent resistance R_{eq}^{123} with

$$R_{eq}^{123} = R_{eq}^{23} + R_1 \tag{6.53}$$

If the voltage between A and C is V, the current I is given by

$$I = \frac{V}{R_{eq}^{23}} = \frac{V}{R_1 + [R_2 R_3/(R_2 + R_3)]}$$

$$= \frac{V(R_2 + R_3)}{R_1 R_2 + R_1 R_3 + R_2 R_3} \tag{6.54}$$

6.11 CELLS, EMF, INTERNAL RESISTANCE

We have already mentioned that a simple device to maintain a steady current in an electric circuit is the electrolytic cell. Basically a cell has two electrodes, called the positive (P0 and the negative (N), as shown in Fig.6.18. They are immersed in an electrolytic solution. Dipped in the solution, the electrodes exchange charges with the electrolyte. The positive electrode has a potential difference V_+ ($V_+ > 0$) between itself and the electrolyte solution immediately adjacent to it marked A in the figure. Similarly, the negative electrode develops a negative potential – (V_-) ($V_- \geq 0$) relative to the electrolyte adjacent to it., marked as B in the figure. When there is no current, the electrolyte has the same potential throughout, so that the potential difference between P and N is $V_+ - (-V_-) = V_+ + V_-$. This difference is called the *electromotive force* (emf) of the cell and is denoted by ε. Thus

$$\varepsilon = V_+ + V_- > 0 \qquad (6.55)$$

Note that ε is, actually, a potential difference and *not a force*. The name emf, however, is used because of historical reasons, and was given at a time when the phenomenon was not understood property.

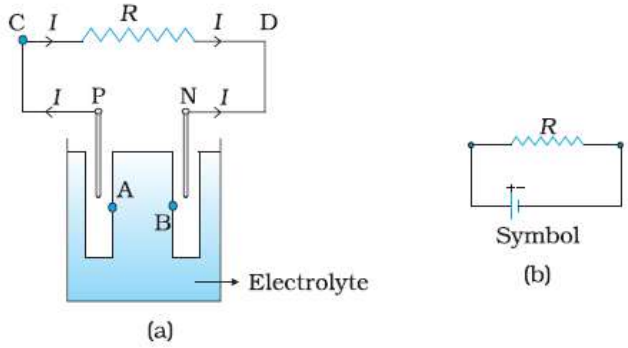


Fig. 6.18 (a) Sketch of an electrolyte cell with positive terminal P and negative terminal N. The gap between the electrodes is exaggerated for clarity. A and B are points in the electrolyte typically close to P and N. (b) the symbol or a cell, + referring to P and – referring to the N electrode. Electrical connections to the cell are made at P and N.

To understand the significance of ε, consider a resistor R connected across the cell (Fig.6.18). A current I flows across R from C to D. As explained before, a steady current is maintained because the electrolyte the same current flows through the electrolyte but from N to P, whereas through R, it flows from P to N.

The electrolyte through which a current flows has a finite resistance r, called the *internal resistance*. Consider first the potential difference between P and N.

Now V = Potential difference between P and A + Potential difference between A and B
+ Potential difference between B and N

$$= \varepsilon \qquad (6.56)$$

Thus, emf ε is the potential difference between the positive and negative electrodes in an open circuit, i.e., when no current is flowing through the cell.

If however R is finite, I is not zero. In that case the potential difference between P and N is

$$V = V_+ + V_- - Ir$$
$$= \varepsilon - Ir \qquad (6.57)$$

Note the negative sign in the expression (*Ir*) for the potential difference between A and B. This is because the current I flows from B and A in the electrolyte.

In practical calculations, internal resistances of cells in the circuit may be neglected when the current I is such that $\varepsilon \gg Ir$. The actual values of the internal resistances of cells vary from cell to cell. The internal resistance of dry cells, however, is much higher than the common electrolytic cells.

We also observe that since V is the potential difference across R, we have from Ohm's law

$$V = I\,R \tag{6.58}$$

Combining Eqs. (6.57) and (6.58), we get

$$I\,R = \varepsilon - I\,r$$

$$\text{Or,} \quad I = \frac{\varepsilon}{R+r} \tag{6.59}$$

The maximum current that can be drawn from a cell is for R = 0 and it is $I_{max} = \varepsilon/r$. However, in most cells the maximum allowed current is much lower than this to prevent permanent damage to the cell.

6.12 CELLS IN SERIES AND IN PARALLEL

Like resistors, cells can be combined together in an electric circuit and like resistors, one can, for calculating currents and voltages in a circuit, replace a combination of cells by an equivalent cell.

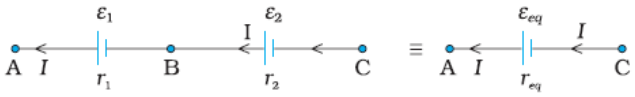


Fig. 6.19 Two cells of emf's ε_1 and ε_2 in the series. $r_1, and\ r_2$ are their internal resistances. For connections across A and C, the combination can be considered as one cell of emf ε_{eq} and an internal resistance r_{eq}.

Consider first two cells in series (Fig.6.20), where one terminal of the two cells is joined together leaving the other terminal in either cell free. $\varepsilon_1, \varepsilon_2$ are emf's of the two cells and r_1, r_2 their internal resistances, respectively.

Let V(A) , V(B), V(C) be the potentials at points A, B and C shown in Fig.6.20. Then V(A) – V(B) is the potential difference between the positive and negative terminals of the first cell. We have already calculated it in Eq.(6.57) and hence,

$$V_{AB} \equiv V(A) - V(B) = \varepsilon_1 - Ir_1 \tag{6.60}$$

Similarly,

$$V_{Bc} \equiv V(B) - V(C) = \varepsilon_2 - Ir_2 \tag{6.61}$$

Hence, the potential difference between the terminals A and C of the combination is

$$V_{AC} \equiv V(A) - V(C) = V(A) - V(B) + V(B) - V(C)$$

$$= (\varepsilon_1 + \varepsilon_2) - I(r_1 + r_2) \tag{6.62}$$

If we wish to replace the combination by a single cell between A and C of emf ε_{eq} and internal resistance r_{eq}, we would have

$$V_{AC} = \varepsilon_{eq} - Ir_{eq} \tag{6.63}$$

Comparing the last two equations, we get

$$\varepsilon_{eq} = \varepsilon_1 + \varepsilon_2 \tag{6.64}$$

$$\text{and, } r_{eq} = r_1 + r_2 \tag{6.65}$$

In Fig 6.19, we had connected the negative electrode of the first to the positive electrode of the second. If instead we connect the two negatives, Eq. (6.61) would change to $V_{Bc} = -\varepsilon_2 - Ir_2$ and we will get

$$\varepsilon_{eq} = \varepsilon_1 - \varepsilon_2 \quad (\varepsilon_1 > \varepsilon_2) \tag{6.66}$$

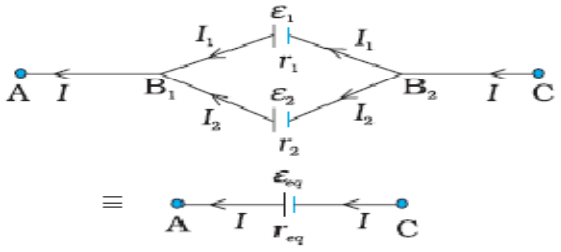


Fig. 6.20 Two cells in parallel. For connections across A and C, the combination can be replaced by one cell of emf ε_{eq} and internal resistances r_{eq} whose values are given in Eqs. (3.64) and (3.65).

The rule for series combination clearly can be extended to any number of cells :

(i) The equivalent emf of a series combination of n cells is just the sum of their individual emf's and

(ii) The equivalent internal resistance of a series combination of n cells is just the sum of their internal resistances.

This is so, when the current leaves each cell from the positive electrode. If in the combination, the current leaves any cell from the *negative* electrode, the emf of the cell enters the expression for ε_{eq} with a *negative* sign, as in Eq.(6.66).

Next, consider a parallel combination of the cells (Fig.6.20). I_1 and I_2 are the currents leaving the positive electrodes of the cells. At the point B_1, I_1 and I_2 flow in whereas the current I flows out. Since as much charge flows in as out, we have

$$I = I_1 + I_2 \tag{6.67}$$

Let $V(B_1)$ and $V(B_2)$ be the potentials at B_1 and B_2 , respectively. Then, considering the first cell, the potential difference across its terminals is $V(B_1) - V(B_2)$. Hence, from Eq.(6.57)

$$V \equiv V(B_1) - V(B_2) = \varepsilon_1 - I_1 r_1 \tag{6.68}$$

Points B_1 and B_2 are connected exactly similarly to the second cell. Hence considering the second cell, we also have

$$V \equiv V(B_1) - V(B_2) = \varepsilon_2 - I_2 r_2 \tag{6.69}$$

Combining the last three equations

$$I = I_1 + I_2$$

$$= \frac{\varepsilon_1 - V}{r_1} + \frac{\varepsilon_2 - V}{r_2} = \left(\frac{\varepsilon_1}{r_1} + \frac{\varepsilon_2}{r_2}\right) - V\left(\frac{1}{r_1} + \frac{1}{r_2}\right) \tag{6.70}$$

Hence, V is given by,

$$V = \frac{\varepsilon_1 r_2 \mp \varepsilon_2 r_1}{r_1 + r_2} - I\,\frac{r_1 r_2}{r_1 + r_2} \tag{6.71}$$

If we want to replace the combination by a single cell, between B_1 and B_2, of emf ε_{eq} and internal resistance r_{eq}, we would have

$$V = \varepsilon_{eq} - I\, r_{eq} \tag{6.72}$$

The last two equations should be the same and hence

$$\varepsilon_{eq} = \frac{\varepsilon_1 r_2 \mp \varepsilon_2 r_1}{r_1 + r_2} \quad (6.73)$$

$$r_{eq} = \frac{r_1 r_2}{r_1 + r_2} \quad (6.74)$$

We can put these equations in a simpler way,

$$\frac{1}{r_{eq}} = \frac{1}{r_1} + \frac{1}{r_2} \quad (6.75)$$

$$\frac{\varepsilon_{eq}}{r_{eq}} = \frac{\varepsilon_1}{r_1} + \frac{\varepsilon_2}{r_2} \quad (6.76)$$

In Fig. 6.20, we had joined the positive terminals together and similarly the two negative ones, so that the currents I_1, I_2 flow out of positive terminals. If the negative terminal of the second is connected to positive terminal of the first, Eqs.(6.75) and (6.76) would still be valid with $\varepsilon_2 \rightarrow -\varepsilon_2$

Gustav Robert Kirchhoff (1824 – 1887)
German physicist, professor at Heidelberg and at Berlin. Mainly known for his development of spectroscopy, he also made many important contributions to mathematical physics, among them, his first and second rules for circuits

.

Equations (6.75) and (6.76) can be extended easily. If there are n cells of emf $\varepsilon_1 \ldots\ldots\ldots \varepsilon_n$ and of internal resistance $r_1 \ldots\ldots\ldots r_n$ respectively, connected in parallel, the combination is equivalent to a single cell of emf ε_{eq} and internal resistance r_{eq} such that

$$\frac{1}{r_{eq}} = \frac{1}{r_1} + \cdots\ldots\ldots + \frac{1}{r_n} \quad (6.77)$$

$$\frac{\varepsilon_{eq}}{r_{eq}} = \frac{\varepsilon_1}{r_1} + \cdots\ldots\ldots + \frac{\varepsilon_n}{r_n} \quad (6.78)$$

6.13 KIRCHHOFF'S RULES

Electric circuit generally consist of a number of resistors and cells inter connected sometimes in a complicated way. The formulae we have derived earlier for series and parallel combinations of resistors are not always sufficient to determine all the currents and potential differences in the circuit. Two rules, called *Kirchhoff's* rules, are very useful for analysis of electric circuits.

Given a circuit , we start by labelling currents in each resistor by a symbol, say I, and a directed arrow to indicate that a current I flows along the resistor in the direction indicated. If ultimately I is determined to be positive, the actual current in the resistor is in the direction of the arrow. If I turns out to be negative, the current actually flows in a direction opposite to the arrow. Similarly, for each source (i.e., cell or some other source of electrical power) the positive and negative electrodes are labelled, as well as, a directed arrow with symbol for the current flowing through the cell. This will tell us the potential difference, V=V(P) – V (N) = ε - I r [Eq.(6.57) between the positive terminal P and the negative terminal N; I here is the current flowing from N to P through the cell]. If, while labelling the current I through the cell one goes from P to N.

Then ofcourse

$$V = \varepsilon + I\,r \quad (6.79)$$

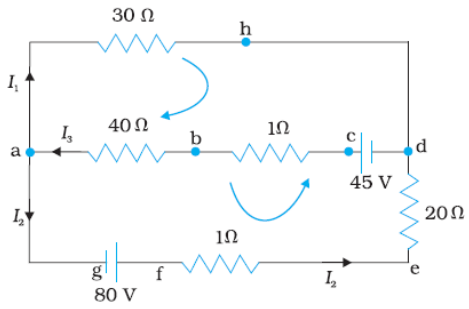


Fig. 6.21 At junction a the current leaving is $I_1 + I_2$ and current entering is I_3. The junction rule says $I_3 = I_1 + I_2$. At point h current entering is I_1. There is only one current leaving h and by junction rule that will also be I_1. For the loops 'ahdcba' and 'ahdefga', the loop rules give $-30I_1 - 41\ I_3 + 45 = 0$ and $-30I_1 + 21\ I_2 - 80 = 0$.

Having clarified labelling, we now state the rules and the proof:

(a) Junction rule: At any junction, the sum of the currents entering the junction is equal to the sum of currents leaving the junction (Fig. 6.21)

This applies equally well if instead of a junction of several lines, we consider a point in a line.

The proof of this rule follows from the fact that when currents are steady, there is no accumulation of charges at any junction or at any point in a line. Thus, the total current flowing in, (which is the rate at which charge flows into the junction), must equal the total current flowing out.

(b) Loop rule: The algebraic sum of changes in potential around any closed loop involving resistors and cells in the loop is zero (Fig. 6.21).

This rule is also obvious, since electric potential is dependent on the location of the point. Thus starting with any point if we come back to the same point, the total change must be zero. In a closed loop, we do come back to the starting point and hence the rule.

It should be noted that because of the symmetry of the network, the great power of Kirchhoff's rules has not been very apparent. In a general network, there will be no such simplification due to symmetry, and by application of Kirchhoff's rules to junctions and closed loops {as many as necessary to solve the unknowns in the network} can we handle the problem.

6.14 WHEATSTONE BRIDGE

As an application of Kirchhoff's rules consider the circuit shown in Fig. 6.22, which is called the *Wheatstone bridge*. The bridge has four resistors R_1, R_2, R_3 and R_4. Across one pair of diagonally opposite points (A and C in the figure) a source is connected. This (i.e., AC) is called the battery arm. Between the other two vertices, B and D, a galvanometer G (which is a device to detect currents) is connected. This line, shown as BD in the figure, is called the galvanometer arm.

For simplicity, we assume that the cell has no internal resistance. In general there will be currents flowing across all the resistors as well as a current I_g through G. Of special interest, is the case of a *balanced* bridge where the resistors are such that $I_g = 0$. We can easily get the balance condition, such that there is no current through G. In this case the Kirchhoff's junction rule applied to junctions D and B (see the figure)

Immediately gives us the relation $I_1 = I_3$ and $I_2 = I_4$. Next, we apply Kirchhoff's loop rule to closed loops ADBA AND CBDC. The first loop gives

$$-I_1R_1 + 0 + I_2R_2 = 0 \qquad (I_g = 0) \qquad (6.81)$$

and the second loop gives, upon using

$I_3 = I_1$, $I_4 = I_2$

$$I_2R_4 + 0 - I_1R_3 = 0 \qquad (6.82)$$

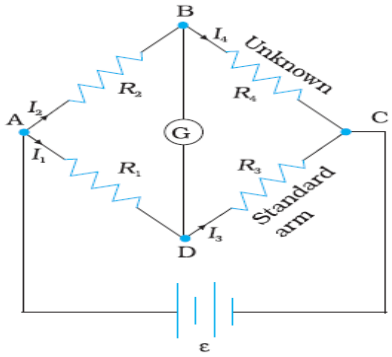


Fig. 6.22

From Eq. 6.81, we obtain

$$\frac{I_1}{I_2} = \frac{R_2}{R_1}$$

whereas from Eq.(6.82), we obtain,

$$\frac{I_1}{I_2} = \frac{R_4}{R_3}$$

Hence, we obtain the condition

$$\frac{R_2}{R_1} = \frac{R_4}{R_3} \qquad [6.83(a)]$$

This last equation relating the four resistors is called the *balance condition* for the galvanometer to give zero or null deflection.

The Wheatstone bridge and its balance condition provide a practical method for determination of an unknown resistance. Let us suppose we have an unknown resistance, which we insert in the fourth arm: R_4 is thus not known. Keeping known resistances R_1 and R_2 in the first and second arm of the bridge, we go on varying R_3 till the galvanometer shows a null deflection. The bridge then is balanced, and from the balance condition the value of the unknown resistance R_4 is given by,

$$R_4 = R_3 \frac{R_2}{R_1} \qquad [6.83(b)]$$

A practical device using this principle is called the *meter bridge*. It will be discussed in the section.

6.15 METER BRIDGE

The meter bridge is shown in Fig 6.23. It consists of a wire of length 1 m and of uniform cross sectional area stretched taut and clamped between two thick metallic strips bent at right angles, as shown. The metallic strip has two gaps across which resistors can be connected. The end points where the wire is clamped are connected to a cell through a key. One end of a galvanometer is connected to the metallic strip midway between the two gaps. The other end of the galvanometer is connected to a 'jockey'. The jockey is essentially a metallic rod whose one end has a knife-edge which can slide over the wire to make electrical connection.

R is an unknown resistance whose value we want to determine. It is connected across one of the gaps. Across the other gap, we connect a standard known resistance S. The jockey is connected to some point D on the wire, a distance *l* cm from the end A. The jockey can be moved along the wire. The portion AD of the wire has a resistance $R_{cm}l$, where R_{cm} is the

resistance of the wire per unit centimetre. The portion DC of the wire similarly has a resistance $R_{cm}(100\text{-}l)$.

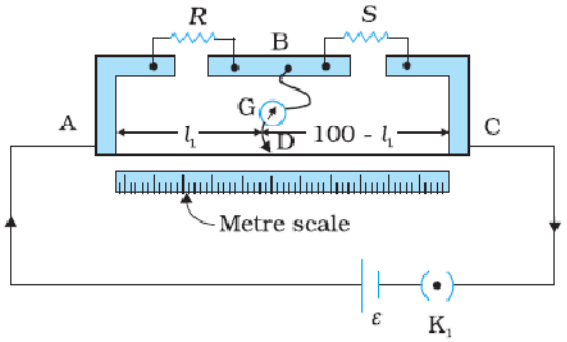


Fig. 6.23 A meter bridge. Wire AC is 1 m long. *R* is a resistance to be measured and *S* is a standard resistance.

The four arms AB, BC, DA and CD [with resistances R, S, $R_{cm}l$ and $R_{cm}(100\text{-}l)$] obviously form a Wheatstone bridge with AC as the battery arm and BD the galvanometer arm. If the jockey is moved along the wire, then there will be one position where the galvanometer will show no current. Let the distance of the jockey from the end A at the balance point be $l = l_1$. The four resistances of the bridge at the balance point then are R, S, R_{cm} l_1 and R_{cm} $(100 - l_1)$. The balance condition, Eq.[6.83(a)] gives

$$\frac{R}{S} = \frac{R_{cm}l_1}{R_{cm}(100-l_1)} = \frac{l_1}{100-l_1} \qquad (6.84)$$

Thus, once we have found out l_1 , the unknown resistance R is known in terms of the standard known resistance S by

$$R = S\,\frac{l_1}{100-l_1} \qquad (6.85)$$

By choosing various values of S, we would get various values of l_1, and calculate R each time. An error in measurement of l_1 would naturally result in an error in R. It can be shown that the percentage error in R can be minimised by adjusting the balance point near the middle of the bridge, i.e., when l_1 is close to 50 cm. (This requires a suitable choice of S.)

6.16 POTENTIOMETER

This is a versatile instrument. It is basically a long piece of uniform wire, sometimes a few meters in length across which a standard cell (B) is connected. In actual design, the wire is sometimes cut in several pieces placed side by side and connected at the ends by thick metal strip (Fig. 6.24). In The figure, the wires run from A to C. The small vertical portions are the thick metal strips connecting the various sections of the wire.

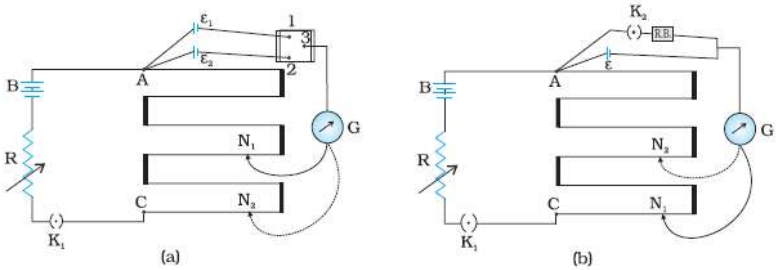


Fig. 6.24 A potentiometer. *G* is a galvanometer and *R* a variable resistance (rheostat). 1, 2, 3 are terminals of a two way key (a) circuit for comparing emf's of two cells; (b) circuit for determining internal resistance of a cell.

A current I flows through the wire which can be varied by a variable resistance (rheostat, R) in the circuit. Since the wire is uniform, the potential difference between A and any point at a distance *l* from A is

$$\varepsilon(l) = \emptyset l \qquad (6.86)$$

where $\emptyset$ is the potential drop per unit length.

Fig. 6.24(a) shows an application of the potentiometer to compare the emf of two cells of emf ε_1 and ε_2 . The points marked 1, 2, 3 form a two way key. Consider first a position of the key where 1 and 3 are connected so that the galvanometer is connected to ε_1 . The jockey is moved along the wire till at a point N_1 , at a distance l_1 from A, there is no deflection in the galvanometer. We can apply Kirchhoff's loop reule to the closed loop AN_1G31A and get

$$\emptyset l_1 + 0 - \varepsilon_1 = 0 \qquad (6.87)$$

Similarly, if another emf ε_2 is balanced against l_2 (AN_2)

$$\emptyset l_2 + 0 - \varepsilon_2 = 0 \qquad (6.88)$$

From the last two equations

$$\frac{\varepsilon_1}{\varepsilon_2} = \frac{l_1}{l_2} \qquad (6.89)$$

This simple mechanism thus allows one to compare the emf's of any two source $(\varepsilon_1, \varepsilon_2)$ .In practice one of the cells is chosen as a standard cell whose emf is known to a high degree of accuracy. The emf of the other cell is then easily calculated from Eq.(6.89).

We can also use a potentiometer to measure internal resistance of a cell [Fig.6.24(b)]. For this the cell (emf ε) whose internal resistance (r) is to be determined is connected across a resistance box through a key K_2 , as shown in the figure . With key K_2 open, balance is obtained at length l_1(AN_1). Then,

$$\varepsilon = \emptyset\, l_1 \qquad [6.90(a)]$$

When key K_2 is closed, the cell sends a current (I) through the resistance box (R). If V is the terminal potential difference of the cell and balance is obtained at length l_2(AN_2),

$$V = \emptyset\, l_2 \qquad [6.90(b)]$$

So, we have $\varepsilon / V = l_1 / l_2$ [6.91(a)]

But, ε = I (r + R) and V = IR . This gives

$$\frac{\varepsilon}{V} = (r + R)/R \qquad [6.92(b)]$$

From eq.[3.94(a)] and [3.94(b)] we have

$$(R + r)/R = l_1 / l_2$$

$$r = R\left(\frac{l_1}{l_2} - 1\right) \qquad (6.93)$$

Using Eq. (6.93) we can find the internal resistance of a given cell.

The potentiometer has the advantage that it draws *no current* from the voltage source being measured. As such it is unaffected by the internal resistance of the source.

SUMMARY

1. *Current* through a given area of a conductor is the net charge passing per unit time through the area.
2. To maintain a steady current, we must have a closed circuit in which an external agency moves electric charge from lower to higher potential energy. The work done per unit charge by the source in taking the charge from lower to higher potential energy (i.e., from one terminal of the source to the other) is called the electromotive force, or *emf*, of the source.

   Note that the emf is not a force; it is the voltage difference between the two terminals of a source in open circuit.
3. *Ohm's law*: The electric current I flowing through a substance is proportional to the voltage V across its ends, i.e., $V \propto I$ or $V = RI$, where R is called the *resistance* of the substance.

   The unit of resistance is ohm: $1\Omega = 1\ V\ A^{-1}$.
4. The *resistance* R of a conductor depends on its length l and constant cross-sectional area A through the relation,

$$R = \frac{\rho l}{A}$$

   where ρ, called *resistivity* is a property of the material and depends on temperature and pressure.
5. *Electrical resistivity* of substances varies over a very wide range. Metals have low resistivity, in the range of 10^{-8} Ω m to 10^{-6} Ω m. Insulators like glass and rubber have 10^{22} to 10^{24} times greater resistivity. Semiconductors like Si and Ge lie roughly in the middle range of resistivity on a logarithmic scale.
6. In most substances, the carriers of current are electrons; in some cases, for example, ionic crystals and electrolytic liquids, positive and negative ions carry the electric current.
7. *Current density* $\mathbf{j}$ gives the amount of charge flowing per second per unit area normal to the flow, $\mathbf{j} = nq\,\mathbf{v}_d$

   where n is the number density (number per unit volume) of charge carriers each of charge q, and $\mathbf{v}_d$ is the *drift velocity* of the charge carriers. For electrons $q = -e$. If $\mathbf{j}$ is normal to a cross-sectional area $\mathbf{A}$ and is constant over the area, the magnitude of the current I through the area is nev_dA.
8. Using $E = V/l$, $I = nev_dA$, and Ohm's law, one obtains

$$\frac{eE}{m} = \rho \frac{ne^2}{m} v_d$$

   The proportionality between the force eE on the electrons in a metal due to the external field E and the drift velocity v_d (not acceleration) can be understood, if we assume that the electrons suffer collisions with ions in the metal, which deflect them randomly. If such collisions occur on an average at a time interval τ,

   $v_d = a\tau = eE\tau/m$

   where a is the acceleration of the electron. This gives

$$\rho = \frac{m}{ne^2\tau}$$

9. In the temperature range in which resistivity increases linearly with temperature, the *temperature coefficient of resistivity* α is defined as the fractional increase in resistivity per unit increase in temperature.

10. Ohm's law is obeyed by many substances, but it is not a fundamental law of nature. It fails if

    (a) *V* depends on *I* non-linearly.

(b) the relation between *V* and *I* depends on the sign of *V* for the same absolute value of *V*.

(c) The relation between *V* and *I* is non-unique.

    An example of (a) is when ρ increases with *I* (even if temperature is kept fixed). A rectifier combines features (a) and (b). GaAs shows the feature (c).

11. When a source of emf ε is connected to an external resistance *R*, the voltage V_{ext} across *R* is given by

$$V_{ext} = IR = \frac{\varepsilon}{R+r} R$$

    where *r* is the *internal resistance* of the source.

12. (a) Total resistance *R* of *n* resistors connected in *series* is given by

    $R = R_1 + R_2 + ... + R_n$

    (b) Total resistance *R* of *n* resistors connected in *parallel* is given by

13. *Kirchhoff's Rules* –

    (a) *Junction Rule:* At any junction of circuit elements, the sum of currents entering the junction must equal the sum of currents leaving it.

    (b) *Loop Rule:* The algebraic sum of changes in potential around any closed loop must be zero.

14. The *Wheatstone bridge* is an arrangement of four resistances – R_1, R_2, R_3, R_4 as shown in the text. The null-point condition is given by

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}$$

    using which the value of one resistance can be determined, knowing the other three resistances.

15. The *potentiometer* is a device to compare potential differences. Since the method involves a condition of *no* current flow, the device can be used to measure potential difference; internal resistance of a cell and compare emf's of two sources.

VERY SHORT ANSWER QUESTIONS (2 MARKS)

1. Define mean free path of electron in a conductor.
2. State Ohm's law and write its mathematical form.
3. Define resistivity or specific resistance.
4. Under what conditions is the current through the mixed grouping of cells maximum ?
5. Why is manganin used for making standard resistors ?
6. The sequence of bands marked on a carbon resistor are : Red, Red, Red , Silver. What is its resistance and tolerance ?
7. Write the color code of a carbon resistor of resistance 23 kilo ohms.
8. Why are household appliances connected in parallel ?

SHORT ANSWER QUESTIONS (4 MARKS)

1. Draw a circuit diagram showing how a potentiometer may be used to find internal resistance of a cell and establish a formula for it.
2. Derive an expression for the effective resistance when three resistors are connected in (i) series (ii) parallel
3. Define electric resistance and write it's SI unit. How does the resistance of a conductor vary if
   (a) Conductor is stretched to 4 times of it's length.
   (b) Temperature of conductor is increased ?
4. Show the variation of current versus voltage graph for GaAs and mark the (i) Non – linear region (ii) Negative resistance region
5. Three identical resistors are connected in parallel and total resistance of the circuit is R/3. Find the value of each resistance.

LONG ANSWER QUESTIONS (8 MARKS)

1. State Kirchhoff's law for an electrical network. Using these laws deduce the condition for balance in a Wheatstone bridge.
2. State the working principle of potentiometer explain with the help of circuit diagram how the emf of two primary cells are compared by using the potentiometer.
3. State the working principle of potentiometer explain with the help of circuit diagram how the potentiometer is used to determine the internal resistance of the given primary cell.

CHAPTER 7

MOVING CHARGES AND MAGNETISM

7.1 INTRODUCTION

Both Electricity and Magnetism have been known for more than 2000 years. However, it was only about 200 years ago, in 1820, that it was realised that they were intimately related*****. During a lecture demonstration in the summer of 1820, Danish physicist Hans Christian Oersted noticed that a current in a straight wire caused a noticeable deflection in a nearby magnetic compass needle. He investigated this phenomenon. He found that the alignment of the needle is tangential to an imaginary circle which has the straight wire as its centre and has its plane perpendicular to the wire. This situation is depicted in Fig.7.1(a). It is noticeable when the current is large and the needle sufficiently closes to the wire so that the earth's magnetic field may be ignored. Reversing the direction of the current reverses the orientation of the needle [Fig. 7.1(b)]. The deflection increases on increasing the current or bringing the needle closer to the wire. Iron filings sprinkled around the wire arrange themselves in concentric circles with the wire as the centre [Fig. 7.1(c)]. Oersted concluded that *moving charges or currents produced a magnetic field in the surrounding space.*

Following this, there was intense experimentation. In 1864, the laws obeyed by electricity and magnetism were unified and formulated by James Maxwell who then realised that light was electromagnetic waves. Radio waves were discovered by Hertz, and produced by J.C.Bose and G. Marconi by the end of the 19th century. A remarkable scientific and technological progress took place in the 20th century. This was due to our increased understanding of electromagnetism and the invention of devices for production, amplification, transmission and detection of electromagnetic waves.

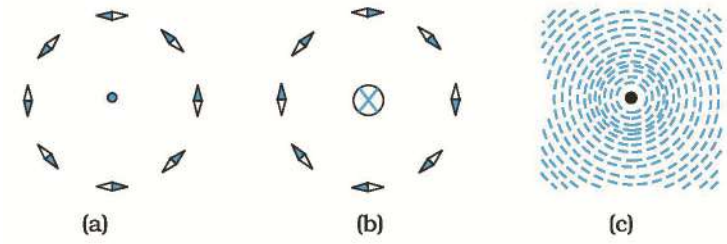


Fig. 7.1 The magnetic field due to a straight long current-carrying wire. The wire is perpendicular to the plane of the paper. A ring of compass needles surrounds the wire. The orientation of the needles is shown when (a) the current emerges out of the plane of the paper, (b) the current moves into the plane of the paper. (c) The arrangement of iron filings around the wire. The darkened ends of the needle represent north poles. The effect of the earth's magnetic field is neglected.

In this chapter, we will see how magnetic field exerts forces on moving charged particles, like electrons, protons, and current-carrying wires. We shall also learn how currents produce magnetic fields. We shall see how particles can be accelerated to very high energies in a cyclotron. We shall study how currents and voltages are detected by a galvanometer.

HANS CHRISTIAN OERSTED (1777–1851)

Hans Christian Oersted (1777–1851) Danish physicist and chemist, professor at Copenhagen. He observed that a compass needle suffers a deflection when placed near a wire carrying an electric current. This discovery gave the first empirical evidence of a connection between electric and magnetic phenomena.

In this and subsequent Chapter on magnetism, we adopt the following convention: A current or a field (electric or magnetic) emerging out of the plane of the paper is depicted by a dot (⊙). A current or a field going into the plane of the paper is depicted by a cross (⊗)*****. Figures. 7.1(a) and 7.1(b) correspond to these two situations, respectively.

***** A dot appears like the tip of an arrow pointed at you, a cross is like the feathered tail of an arrow moving away from you.

7.2 MAGNETIC FORCE

7.2.1 Sources and fields

Before we introduce the concept of a magnetic field **B**, we shall recapitulate what we have learnt in Chapter 1 about the electric field **E.** We have seen that the interaction between two charges can be considered in two stages. The charge Q, the source of the field, produces an electric field **E**, where

$$\mathbf{E} = Q/(4\pi\varepsilon)r^2 \qquad (7.1)$$

where $\hat{r}$ is unit vector along **r**, and the field **E** is a vector field. A charge q interacts with this field and experiences a force **F** given by

$$\mathbf{F} = q\,\mathbf{E} = q\,Q\,\hat{r}/(4\pi\varepsilon)\,r^2 \qquad (7.2)$$

HENDRIK ANTOON LORENTZ (1853 – 1928)

Hendrik Antoon Lorentz (1853–1928) Dutch theoretical physicist, professor at Leiden. He investigated the relationship between electricity, magnetism, and mechanics. In order to explain the observed effect of magnetic fields on emitters of light (Zeeman effect), he postulated the existence of electric charges in the atom, for which he was awarded the Nobel Prize in 1902. He derived a set of transformation equations (known after him, as Lorentz transformation equations) by some tangled mathematical arguments, but he was not aware that these equations hinge on a new concept of space and time.

As pointed out in the Chapter 1, the field **E** is not just an artefact but has a physical role. It can convey energy and momentum and is not established instantaneously but takes finite time to propagate. The concept of a field was specially stressed by Faraday and was incorporated by Maxwell in his unification of electricity and magnetism. In addition to depending on each point in space, it can also vary with time, i.e., be a function of time. In our discussions in this chapter, we will assume that the fields do not change with time.

The field at a particular point can be due to one or more charges. If there are more charges the fields add vectorially. You have already learnt in Chapter 1 that this is called the principle of superposition. Once the field is known, the force on a test charge is given by Eq. (7.2).

Just as static charges produce an electric field, the currents or moving charges produce (in addition) a magnetic field, denoted by **B (r)**, again a vector field. It has several basic properties identical to the electric field. It is defined at each point in space (and can in addition depend on time). Experimentally, it is found to obey the principle of superposition: *the magnetic field of several sources is the vector addition of magnetic field of each individual source.*

7.2.2 Magnetic Field, Lorentz Force

Let us suppose that there is a point charge q (moving with a velocity **v** and, located at **r** at a given time t) in presence of both the electric field **E (r)** and the magnetic field **B (r)**. The force on an electric charge q due to both of them can be written as

$$\mathbf{F} = q\,[\,\mathbf{E}(\mathbf{r}) + \mathbf{v} \times \mathbf{B}(\mathbf{r})] \equiv \mathbf{F}_{\text{electric}} + \mathbf{F}_{\text{magnetic}} \qquad (7.3)$$

This force was given first by H.A. Lorentz based on the extensive experiments of Ampere and others. It is called the *Lorentz force*. You have already studied in detail the force due to the electric field. If we look at the interaction with the magnetic field, we find the following features.

(i) It depends on q, **v** and **B** (charge of the particle, the velocity and the magnetic field). *Force on a negative charge is opposite to that on a positive charge.*

(ii) The magnetic force q [**v** × **B**] includes a vector product of velocity and magnetic field. The vector product makes the force due to magnetic field vanish (become zero) if velocity and magnetic field are parallel or anti-parallel. The force acts in a (sideways) direction perpendicular to both the velocity and the magnetic field. Its direction is given by the screw rule or right hand rule for vector (or cross) product as illustrated in Fig. 7.2.

(iii) The magnetic force is zero if charge is not moving (as then |**v**|= 0). Only a moving charge feels the magnetic force.

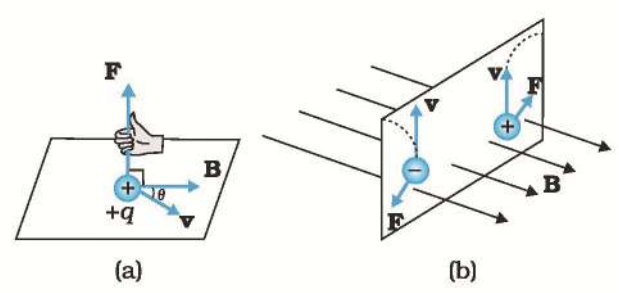


Fig. 7.2 The direction of the magnetic force acting on a charged particle. (a) The force on a positively charged particle with velocity **v** and making an angle θ with the magnetic field **B** is given by the right-hand rule. (b) A moving charged particle q is deflected in an opposite sense to $-q$ in the presence of magnetic field.

The expression for the magnetic force helps us to define the unit of the magnetic field, if one takes q, **F** and **v**, all to be unity in the force equation $\mathbf{F} = q\,[\,\mathbf{v} \times \mathbf{B}] = q\,v\,B \sin\theta\,\hat{n}$, where θ is the angle between **v** and **B** [see Fig. 7.2 (a)]. The magnitude of magnetic field B is 1 SI unit, when the force acting on a unit charge (1 C), moving perpendicular to **B** with a speed 1m/s, is one newton.

Dimensionally, we have $[B] = [F/qv]$ and the unit of **B** are Newton second / (coulomb metre). This unit is called *tesla* (T) named after Nikola Tesla (1856 – 1943). Tesla is a rather large unit. A smaller unit (non-SI) called *gauss* ($=10^{-4}$ tesla) is also often used. The earth's magnetic field is about 3.6×10^{-5} T. Table 7.1 lists magnetic fields over a wide range in the universe.

TABLE 7.1 ORDER OF MAGNITUDES OF MAGNETIC FIELDS IN A VARIETY OF PHYSICAL SITUATIONS

Physical situation	Magnitude of B (in tesla)
Surface of a neutron star	10^{8}
Typical large field in a laboratory	1
Near a small bar magnet	10^{-2}
On the earth's surface	10^{-6}
Human nerve fibre	10^{-10}
Interstellar space	10^{-12}

7.2.3 Magnetic force on a current-carrying conductor

We can extend the analysis for force due to magnetic field on a single moving charge to a straight rod carrying current. Consider a rod of a uniform cross-sectional area A and length l. We shall assume one kind of mobile carriers as in a conductor (here electrons). Let the number density of these mobile charge carriers in it be n. Then the total number of mobile charge carriers in it is nlA. For a steady current I in this conducting rod, we may assume that each mobile carrier has an average drift velocity $\mathbf{v}d$ (see Chapter 6). In the presence of an external magnetic field $\mathbf{B}$, the force on these carriers is:

$$\mathbf{F} = (nlA)q\ \mathbf{v}_d \times \mathbf{B}$$

where q is the value of the charge on a carrier. Now $nq\ \mathbf{v}$d is the current density $\mathbf{j}$ and $|(nq\ \mathbf{v}_d)|A$ is the current I (see Chapter 6 for the discussion of current and current density). Thus,

$$\mathbf{F} = [(nq\ \mathbf{v}_d)l\ A] \times \mathbf{B} = [\ \mathbf{j}Al\ ] \times \mathbf{B}$$

$$= I\boldsymbol{l} \times \mathbf{B} \qquad (7.4)$$

where $\boldsymbol{l}$ is a vector of magnitude l, the length of the rod, and with a direction identical to the current I. Note that the current I is not a vector. In the last step leading to Eq. (4.4), we have transferred the vector sign from $\mathbf{j}$ to $\boldsymbol{l}$.

Equation (7.4) holds for a straight rod. In this equation, $\mathbf{B}$ is the external magnetic field. It is not the field produced by the current-carrying rod. If the wire has an arbitrary shape we can calculate the Lorentz force on it by considering it as a collection of linear strips $\mathrm{d}\boldsymbol{l}_\mathrm{j}$ and summing

$$\mathbf{F} = \sum_\mathrm{j} \mathrm{Id}\,\boldsymbol{l}_\mathrm{j} \times \mathbf{B}$$

This summation can be converted to an integral in most cases.

ON PERMITTIVITY AND PERMEABILITY

In the universal law of gravitation, we say that any two point masses exert a force on each other which is proportional to the product of the masses m_1, m_2 and inversely proportional to the square of the distance r between them. We write it as $F = Gm_1m_2/r^2$ where G is the universal constant of gravitation. Similarly, in Coulomb's law of electrostatics we write the force between two point charges q_1, q_2, separated by a distance r as $F = kq_1\ q_2/r^2$ where k is a constant of proportionality. In SI units, k is taken as $1/4\pi\varepsilon$ where ε is the permittivity of the medium. Also in magnetism, we get another constant, which in SI units, is taken as $\mu/4\pi$ where μ is the permeability of the medium.

Although G, ε and μ arise as proportionality constants, there is a difference between gravitational force and electromagnetic force. While the gravitational force does not depend on the intervening medium, the electromagnetic force depends on the medium between the two charges or magnets. Hence, while G is a universal constant, ε and μ depend on the medium. They have different values for different media. The product $\varepsilon\mu$ turns out to be related to the speed v of electromagnetic radiation in the medium through $\varepsilon\mu = 1/v^2$.

Electric permittivity ε is a physical quantity that describes, how an electric field affects and is affected by a medium. It is determined by the ability of a material to polarise in response to an applied field, and thereby to cancel, partially, the field inside the material. Similarly, magnetic permeability μ is the ability of a substance to acquire magnetisation in magnetic fields. It is a measure of the extent to which magnetic field can penetrate matter.

7.3 MOTION IN A MAGNETIC FIELD

We will now consider, in greater detail, the motion of a charge moving in a magnetic field. We have learnt in Mechanics (see Class XI book, Chapter 6) that a force on a particle does work if the force has a component along (or opposed to) the direction of motion of the particle. In the case of motion of a charge in a magnetic field, the magnetic force is perpendicular to the velocity of the particle. So no work is done and no change in the magnitude of the velocity is produced (though the direction of momentum may be changed). [Notice that this is unlike the force due to an electric field, q **E**, which *can* have a component parallel (or antiparallel) to motion and thus can transfer energy in addition to momentum.]

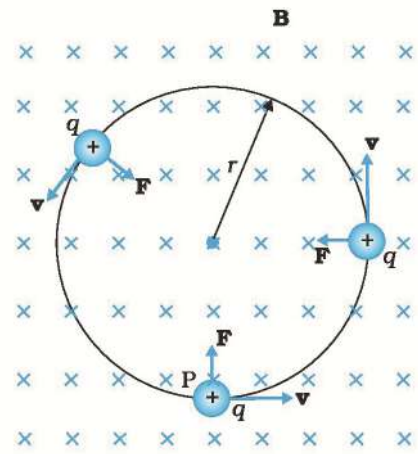


Fig. 7.5 Circular motion

We shall consider motion of a charged particle in a *uniform* magnetic field. First consider the case of **v** perpendicular to **B**. The perpendicular force, q **v** × **B**, acts as a centripetal force and produces a circular motion perpendicular to the magnetic field. *The particle will describe a circle if* **v** *and* **B** *are perpendicular to each other* (Fig. 7.5).

If velocity has a component along **B**, this component remains unchanged as the motion along the magnetic field will not be affected by the magnetic field. The motion in a plane perpendicular to **B** is as before a circular one, thereby producing a *helical motion* (Fig. 7.6).

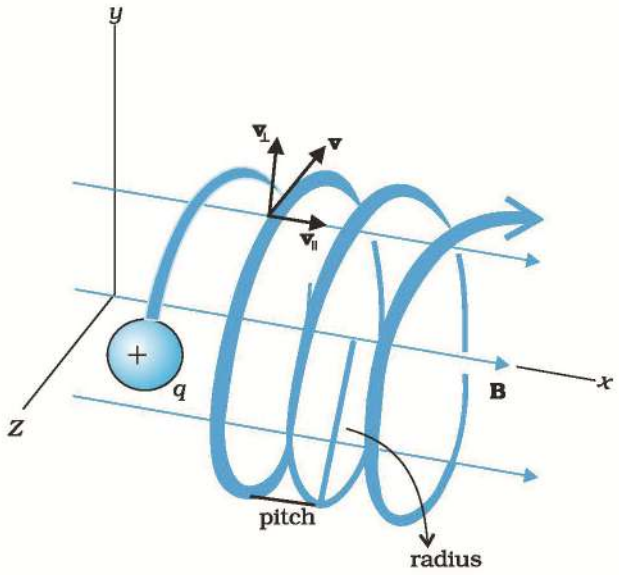


Fig. 7.6 Helical motion

You have already learnt in earlier classes (See Class XI, Chapter 4) that if r is the radius of the circular path of a particle, then a force of $m\,v^2 / r$, acts perpendicular to the path towards the centre of the circle, and is called the centripetal force. If the velocity **v** is perpendicular to the magnetic field **B,** the magnetic force is perpendicular to both **v** and **B** and acts like a centripetal force. It has a magnitude $q\,v\,B$. Equating the two expressions for centripetal force,

$$m\,v^2/r = q\,v\,B, \text{ which gives}$$

$$r = m\,v / qB \qquad (7.5)$$

for the radius of the circle described by the charged particle. The larger the momentum, the larger is the radius and bigger the circle described. If ω is the angular frequency, then $v = \omega r$. So,

$$\omega = 2\pi\nu = q\,B/\,m \qquad [7.6(a)]$$

which is independent of the velocity or energy . Here $\nu$is the frequency of rotation. The independence of $\nu$from energy has important application in the design of a cyclotron (see Section 7.4.2).

The time taken for one revolution is $T = 2\pi/\omega \equiv 1/\nu$. If there is a component of the velocity parallel to the magnetic field (denoted by $v_{\|}$), it will make the particle move along the field and the path of the particle would be a helical one (Fig. 7.6). The distance moved along the magnetic field in one rotation is called pitch p. Using Eq. [7.6 (a)], we have

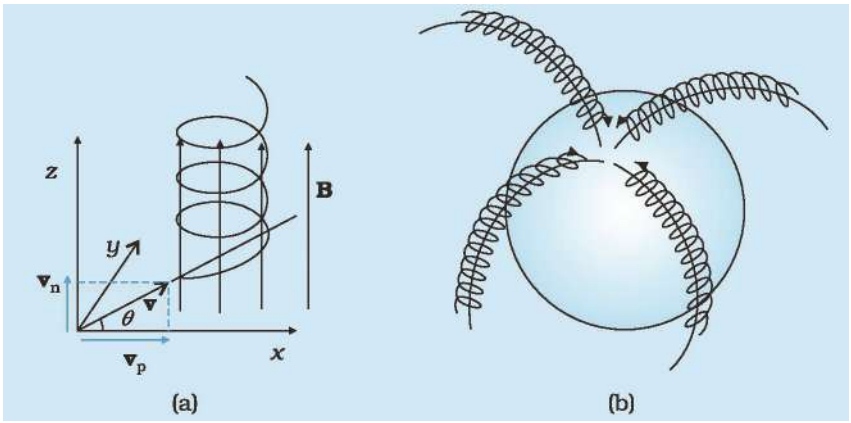
$$p = v_{\|}T = 2\pi m\, v_{\|} / q\, B \qquad [7.6(b)]$$

The radius of the circular component of motion is called the *radius* of the *helix*.

HELICAL MOTION OF CHARGED PARTICLES AND AURORA BOREALIS

In polar regions like Alaska and Northern Canada, a splendid display of colours is seen in the sky. The appearance of dancing green pink lights is fascinating, and equally puzzling. An explanation of this natural phenomenon is now found in physics, in terms of what we have studied here.

Consider a charged particle of mass m and charge q, entering a region of magnetic field **B** with an initial velocity **v**. Let this velocity have a component $\mathbf{v}_p$ parallel to the magnetic field and a component **v**n normal to it. There is no force on a charged particle in the direction of the field. Hence the particle continues to travel with the velocity $\mathbf{v}_p$ parallel to the field. The normal component **v**n of the particle results in a Lorentz force ($\mathbf{v}_n \times \mathbf{B}$) which is perpendicular to both $\mathbf{v}_n$ and **B**. As seen in Section 7.3.1 the particle thus has a tendency to perform a circular motion in a plane perpendicular to the magnetic field. When this is coupled with the velocity parallel to the field, the resulting trajectory will be a helix along the magnetic field line, as shown in Figure (a) here. Even if the field line bends, the helically moving particle is trapped and guided to move around the field line. Since the Lorentz force is normal to the velocity of each point, the field does no work on the particle and the magnitude of velocity remains the same.



During a solar flare, a large number of electrons and protons are ejected from the sun. Some of them get trapped in the earth's magnetic field and move in helical paths along the field lines. The field lines come closer to each other near the magnetic poles; see figure (b). Hence the density of charges increases near the poles. These particles collide with atoms and molecules of the atmosphere. Excited oxygen atoms emit green light and excited nitrogen atoms emits pink light. This phenomenon is called *Aurora Borealis* in physics.

7.4 MOTION IN COMBINED ELECTRIC AND MAGNETIC FIELDS

7.4.1 Velocity selector

You know that a charge q moving with velocity **v** in presence of both electric and magnetic fields experiences a force given by Eq. (7.3), that is,

$\mathbf{F} = q\,(\mathbf{E} + \boldsymbol{v} \times \mathbf{B}) = \mathbf{F}_E + \mathbf{F}_B$

We shall consider the simple case in which electric and magnetic fields are perpendicular to each other and also perpendicular to the velocity of the particle, as shown in Fig. 7.7. We have,

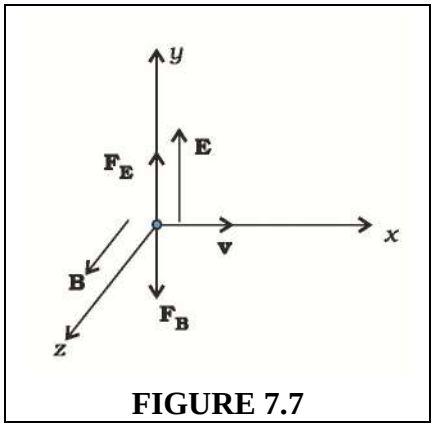


FIGURE 7.7

$\mathbf{E} = E\,\hat{\mathrm{j}}$, $\mathbf{B} = B\,\hat{\mathrm{k}}$, $\mathbf{v} = v\,\hat{\mathrm{i}}$

$\mathbf{F}_E = q\mathbf{E} = qE\,\hat{\mathrm{j}}$, $\mathbf{F}_B = q\mathbf{v} \times \mathbf{B}$, $= q\,(v\,\hat{\mathrm{i}} \times B\,\hat{\mathrm{k}}) = -qB\,\hat{\mathrm{j}}$

Therefore, $\mathbf{F} = q\,(E - vB)\,\hat{\mathrm{j}}$

Thus, electric and magnetic forces are in opposite directions as shown in the figure. Suppose, we adjust the value of **E** and **B** such that magnitudes of the two forces are equal. Then, total force on the charge is zero and the charge will move in the fields undeflected. This happens when,

$$qE = qvB \qquad \text{or } v = \frac{E}{B} \tag{7.7}$$

This condition can be used to select charged particles of a particular velocity out of a beam containing charges moving with different speeds (irrespective of their charge and mass). The crossed *E* and *B* fields, therefore, serve as a *velocity selector*. Only particles with speed *E*/*B* pass undeflected through the region of crossed fields. This method was employed by J. J. Thomson in 1897 to measure the charge to mass ratio (*e*/*m*) of an electron. The principle is also employed in Mass Spectrometer – a device that separates charged particles, usually ions, according to their charge to mass ratio.

7.4.2 Cyclotron

The cyclotron is a machine to accelerate charged particles or ions to high energies. It was invented by E.O. Lawrence and M.S. Livingston in 1934 to investigate nuclear structure. The cyclotron uses both electric and magnetic fields in combination to increase the energy of charged particles. As the fields are perpendicular to each other they are called *crossed fields*. Cyclotron uses the fact that the frequency of revolution of the charged particle in a magnetic field is independent of its energy. The particles move most of the time inside two semicircular disc-like metal containers, D_1 and D_2, which are called *dees* as they look like the letter D. Figure 7.8 shows a schematic view of the cyclotron. Inside the metal boxes the particle is shielded and is not acted on by the electric field. The magnetic field, however, acts on the particle and makes it go round in a circular path inside a dee. Every time the particle moves from one dee to another it is acted upon by the electric field. The sign of the electric field is changed alternately in tune with the circular motion of the particle. This ensures that the particle is always accelerated by the electric field. Each time the acceleration increases the energy of the particle. As energy increases, the radius of the circular path increases. So the path is a spiral one.

The whole assembly is evacuated to minimise collisions between the ions and the air molecules. A high frequency alternating voltage is applied to the dees. In the sketch shown in Fig. 7.8, positive ions or positively charged particles (e.g., protons) are released at the centre P. They move in a semi-circular path in one of the dees and arrive in the gap between the dees in a time interval *T*/2; where *T*, the period of revolution, is given by Eq.(7.6),

$$T = \frac{1}{\nu_c} = \frac{2\pi m}{qB}$$

$$\text{or} \quad \nu_c = \frac{qB}{2\pi m} \tag{7.8}$$

This frequency is called the *cyclotron frequency* for obvious reasons and is denoted by ν_c.

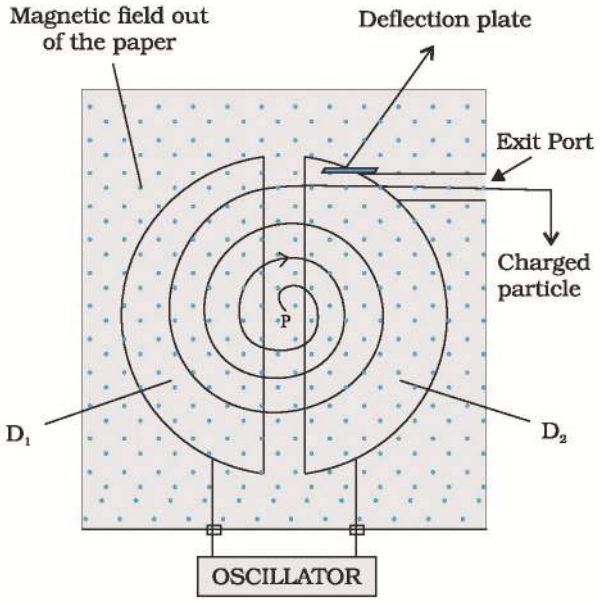


Fig.7.8 A schematic sketch of the cyclotron. There is a source of charged particles or ions at P which move in a circular fashion in the dees, D1 and D2, on account of a uniform perpendicular magnetic field B. An alternating voltage source accelerates these ions to high speeds. The ions are eventually 'extracted' at the exit port.

The frequency *va* of the applied voltage is adjusted so that the polarity of the dees is reversed in the same time that it takes the ions to complete one half of the revolution. The requirement $\nu_a = \nu_c$ is called the *resonance condition*. The phase of the supply is adjusted so that when the positive ions arrive at the edge of D_1, D_2 is at a lower potential and the ions are accelerated across the gap. Inside the dees the particles travel in a region free of the electric field. The increase in their kinetic energy is qV each time they cross from one dee to another (V refers to the voltage across the dees at that time). From Eq. (7.5), it is clear that the radius of their path goes on increasing each time their kinetic energy increases. The ions are repeatedly accelerated across the dees until they have the required energy to have a radius approximately that of the dees. They are then deflected by a magnetic field and leave the system via an exit slit. From Eq. (7.5) we have,

$$v = \frac{qBR}{m} \tag{7.9}$$

where R is the radius of the trajectory at exit, and equals the radius of a dee.

Hence, the kinetic energy of the ions is,

$$\frac{1}{2} m v^2 = \frac{q^2 B^2 R^2}{2m} \tag{7.10}$$

The operation of the cyclotron is based on the fact that the time for one revolution of an ion is independent of its speed or radius of its orbit. The cyclotron is used to bombard nuclei with energetic particles, so accelerated by it, and study the resulting nuclear reactions. It is also used to implant ions into solids and modify their properties or even synthesise new materials. It is used in hospitals to produce radioactive substances which can be used in diagnosis and treatment.

ACCELERATORS IN INDIA

India has been an early entrant in the area of accelerator-based research. The vision of Dr. Meghnath Saha created a 37" Cyclotron in the Saha Institute of Nuclear Physics in Kolkata in 1953. This was soon followed by a series of Cockroft-Walton type of accelerators established in Tata Institute of Fundamental Research (TIFR), Mumbai, Aligarh Muslim University (AMU), Aligarh, Bose Institute, Kolkata and Andhra University, Waltair.

The sixties saw the commissioning of a number of Van de Graaff accelerators: a 5.5 MV terminal machine in Bhabha Atomic Research Centre (BARC), Mumbai (1963); a 2 MV terminal machine in Indian Institute of Technology (IIT), Kanpur; a 400 kV terminal machine in Banaras Hindu University (BHU), Varanasi; and Punjabi University, Patiala. One 66 cm Cyclotron donated by the Rochester University of USA was commissioned in Panjab University, Chandigarh. A small electron accelerator was also established in University of Pune, Pune.

In a major initiative taken in the seventies and eighties, a Variable Energy Cyclotron was built indigenously in Variable Energy Cyclotron Centre (VECC), Kolkata; 2 MV Tandem Van de Graaff accelerator was developed and built in BARC and a 14 MV Tandem Pelletron accelerator was installed in TIFR.

This was soon followed by a 15 MV Tandem Pelletron established by University Grants Commission (UGC), as an inter-university facility in Inter-University Accelerator Centre (IUAC), New Delhi; a 3 MV Tandem Pelletron in Institute of Physics, Bhubaneswar; and two 1.7 MV Tandetrons in Atomic Minerals Directorate for Exploration and Research, Hyderabad and Indira Gandhi Centre for Atomic Research, Kalpakkam. Both TIFR and IUAC are augmenting their facilities with the addition of superconducting LINAC modules to accelerate the ions to higher energies.

Besides these ion accelerators, the Department of Atomic Energy (DAE) has developed many electron accelerators. A 2 GeV Synchrotron Radiation Source is being built in Raja Ramanna Centre for Advanced Technologies, Indore.

The Department of Atomic Energy is considering Accelerator Driven Systems (ADS) for power production and fissile material breeding as future options.

7.5 MAGNETIC FIELD DUE TO A CURRENT ELEMENT, BIOT-SAVART LAW

All magnetic fields that we know are due to currents (or moving charges) and due to intrinsic magnetic moments of particles. Here, we shall study the relation between current and the magnetic field it produces. It is given by the Biot-Savart's law. Figure 7.9 shows a finite conductor XY carrying current *I*. Consider an infinitesimal element d*l* of the conductor. The magnetic field d**B** due to this element is to be determined at a point P which is at a distance *r* from it. Let θ be the angle between d***l*** and the displacement vector **r**. According to Biot-Savart's law, the magnitude of the magnetic field d**B** is proportional to the current *I*, the element length |d***l***|, and inversely proportional to the square of the distance *r*. Its direction***** is perpendicular to the plane containing d***l*** and **r** . Thus, in vector notation,

$$d\mathbf{B} \propto \frac{Id\mathbf{l}\times\mathbf{r}}{r^2}$$

$$= \frac{\mu_0}{4\pi}\frac{Id\mathbf{l}\times\mathbf{r}}{r^3} \qquad [7.11(a)]$$

where $\mu_0/4\pi$ is a constant of proportionality. The above expression holds when the medium is vacuum.

The magnitude of this field is,

$$|d\mathbf{B}| = \frac{\mu_0}{4\pi}\frac{Id\mathbf{l}\sin\theta}{r^2} \qquad [7.11(b)]$$

where we have used the property of cross-product. Equation [7.11 (a)] constitutes our basic equation for the magnetic field. The proportionality constant in SI units has the exact value,

$$\frac{\mu_0}{4\pi} = 10^{-7}\ \text{Tm/A}\ 4\pi \qquad [7.11(c)]$$

We call μ_0 the *permeability* of free space (or vacuum).

The Biot-Savart law for the magnetic field has certain similarities, as well as, differences with the Coulomb's law for the electrostatic field. Some of these are:

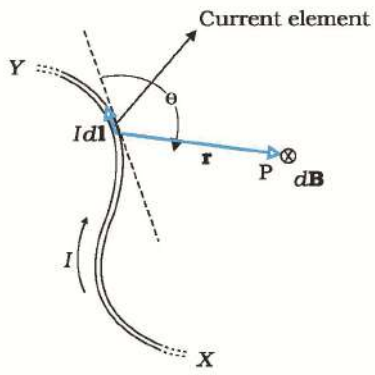


Fig. 7.9 Illustration of the Biot-Savart law. The current element $I\,d\boldsymbol{l}$ produces a field $d\mathbf{B}$ at a distance r. The ⊗ sign indicates that the field is perpendicular to the plane of this page and directed into it.

(i) Both are long range, since both depend inversely on the square of distance from the source to the point of interest. The principle of superposition applies to both fields. [In this connection, note that the magnetic field is *linear* in the *source* $I\,d\boldsymbol{l}$ just as the electrostatic field is linear in its source: the electric charge.]

(ii) The electrostatic field is produced by a scalar source, namely, the electric charge. The magnetic field is produced by a vector source $I\,d\boldsymbol{l}$.

* The sense of $d\boldsymbol{l} \times \mathbf{r}$ is also given by the *Right Hand Screw rule*: Look at the plane containing vectors $d\boldsymbol{l}$ and $\mathbf{r}$. Imagine moving from the first vector towards second vector. If the movement is anticlockwise, the resultant is towards you. If it is clockwise, the resultant is away from you.

(iii) The electrostatic field is along the displacement vector joining the source and the field point. The magnetic field is perpendicular to the plane containing the displacement vector $\mathbf{r}$ and the current element $I\,d\boldsymbol{l}$.

(iv) There is angle dependence in the Biot-Savart law which is not present in the electrostatic case. In Fig. 7.9, the magnetic field at any point in the direction of $d\boldsymbol{l}$ (the dashed line) is zero. Along this line, $\theta = 0$, $\sin\theta = 0$ and from Eq. [7.11(a)], $|d\mathbf{B}| = 0$.

There is an interesting relation between ε_0, the permittivity of free space; μ , the permeability of free space; and c, the speed of light in vacuum :

$$\varepsilon_0\mu_0 = (4\pi\varepsilon_0)\frac{\mu_0}{4\pi} = \frac{1}{9\times10^9}(10^{-7}) = \frac{1}{(3\times10^8)^2} = \frac{1}{c^2}$$

We will discuss this connection further in Chapter 8 on the electromagnetic waves. Since the speed of light in vacuum is constant, the product $\mu_0\varepsilon_0$ is fixed in magnitude. Choosing the value of either ε_0 or μ_0, fixes the value of the other. In SI units, μ_0 is fixed to be equal to $4\pi \times 10^{-7}$ in magnitude.

In the next section, we shall use the Biot-Savart law to calculate the magnetic field due to a circular loop.

7.6 MAGNETIC FIELD ON THE AXIS OF A CIRCULAR CURRENT LOOP

In this section, we shall evaluate the magnetic field due to a circular coil along its axis. The evaluation entails summing up the effect of infinitesimal current elements ($I\,d\boldsymbol{l}$)

mentioned in the previous section. We assume that the current I is steady and that the evaluation is carried out in free space (i.e., vacuum).

Figure 7.11 depicts a circular loop carrying a steady current I. The loop is placed in the y-z plane with its centre at the origin O and has a radius R. The x-axis is the axis of the loop. We wish to calculate the magnetic field at the point P on this axis. Let x be the distance of P from the centre O of the loop.

Consider a conducting element d$\boldsymbol{l}$ of the loop. This is shown in Fig. 7.11. The magnitude dB of the magnetic field due to d$\boldsymbol{l}$ is given by the Biot-Savart law [Eq. 7.11(a)],

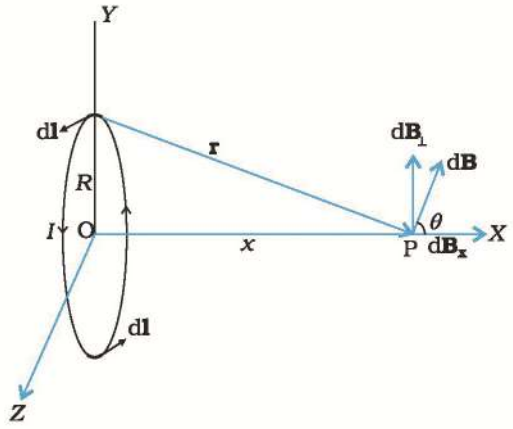


Fig. 7.11 Magnetic field on the axis of a current carrying circular loop of radius R. Shown are the magnetic field $d\mathbf{B}$ (due to a line element $d\boldsymbol{l}$) and its components along and perpendicular to the axis.

$$dB = \frac{\mu_0}{4\pi}\frac{I|d\mathbf{l}\times\mathbf{r}|}{r^3} \tag{7.12}$$

Now $r^2 = x^2 + R^2$. Further, any element of the loop will be perpendicular to the displacement vector from the element to the axial point. For example, the element d$\boldsymbol{l}$ in Fig. 7.11 is in the y-z plane, whereas, the displacement vector $\mathbf{r}$ from d$\boldsymbol{l}$ to the axial point P is in the x-y plane. Hence $|d\boldsymbol{l} \times \mathbf{r}| = r\,dl$. Thus,

$$\mathrm{d}B = \frac{\mu_0}{4\pi}\frac{Id\mathbf{l}}{\left(r^2+R^2\right)} \tag{7.13}$$

The direction of d$\mathbf{B}$ is shown in Fig. 7.11. It is perpendicular to the plane formed by d$\boldsymbol{l}$ and $\mathbf{r}$. It has an x-component d$\mathbf{B}_x$ and a component perpendicular to x-axis, d$\mathbf{B}_\perp$. When the components perpendicular to the x-axis are summed over, they cancel out and we obtain a null result. For example, the d$\mathbf{B}_\perp$ component due to d$\boldsymbol{l}$ is cancelled by the contribution due to the diametrically opposite d$\boldsymbol{l}$ element, shown in Fig. 7.11. Thus, only the x-component survives. The net contribution along x-direction can be obtained by integrating $\mathrm{d}B_x = \mathrm{d}B\cos\theta$ over the loop. For Fig. 7.11,

$$\cos\theta = \frac{R}{\left(x^2+R^2\right)^{1/2}} \tag{7.14}$$

From Eqs. (7.13) and (7.14)

$$\mathrm{d}B_x = \frac{\mu_0 Idl}{4\pi}\frac{R}{\left(x^2+R^2\right)^{3/2}}$$

The summation of elements dl over the loop yields $2\pi R$, the circumference of the loop. Thus, the magnetic field at P due to entire circular loop is

$$\mathbf{B} = B_x\hat{\mathbf{i}} = \frac{\mu_0 IR^2}{2\left(x^2+R^2\right)^{3/2}}\hat{\mathbf{i}} \tag{7.15}$$

As a special case of the above result, we may obtain the field at the centre of the loop. Here $x = 0$, and we obtain,

$$\mathbf{B}_0 = \frac{\mu_0 I}{2R}\hat{\mathbf{i}} \qquad (7.12)$$

The magnetic field lines due to a circular wire form closed loops and are shown in Fig. 7.12. The direction of the magnetic field is given by (another) *right-hand thumb rule* stated below:

Curl the palm of your right hand around the circular wire with the fingers pointing in the direction of the current. The right-hand thumb gives the direction of the magnetic field.

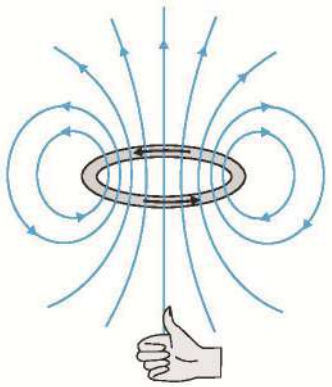


Fig. 7.12 The magnetic field lines for a current loop. The direction of the field is given by the right-hand thumb rule described in the text. The upper side of the loop may be thought of as the north pole and the lower side as the south pole of a magnet.

7.7 AMPERE'S CIRCUITAL LAW

There is an alternative and appealing way in which the Biot-Savart law may be expressed. Ampere's circuital law considers an open surface with a boundary (Fig. 7.14). The surface has current passing through it. We consider the boundary to be made up of a number of small line elements. Consider one such element of length *dl*. We take the value of the tangential component of the magnetic field, B_t, at this element and multiply it by the length of that element *dl* [Note: $B_t dl = \mathbf{B}.d\boldsymbol{l}$]. All such products are added together. We consider the limit as the lengths of elements get smaller and their number gets larger. The sum then tends to an integral. Ampere's law states that this integral is equal to μ_0 times the total current passing through the surface, i.e.,

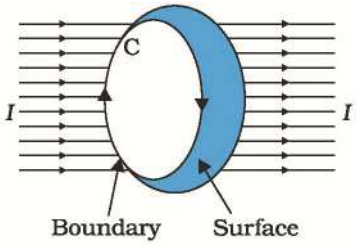


Fig. 7.14

$$\oint \mathbf{B}.d\boldsymbol{l} = \mu_0 I \qquad [7.17(a)]$$

where I is the total current through the surface. The integral is taken over the closed loop coinciding with the boundary C of the surface. The relation above involves a sign-convention, given by the right-hand rule. Let the fingers of the right-hand be curled in the sense the boundary is traversed in the loop integral $\mathbf{B}.d\boldsymbol{l}$. Then the direction of the thumb gives the sense in which the current I is regarded as positive.

For several applications, a much simplified version of Eq. [7.17(a)] proves sufficient. We shall assume that, in such cases, it is possible to choose the loop (called an *amperian loop*) such that at each point of the loop, *either*

(i) **B** is tangential to the loop and is a non-zero *constant* B, *or*

(ii) **B** is normal to the loop, *or*

(iii) **B** vanishes.

Now, let L be the length (part) of the loop for which **B** is tangential**.** Let I_e be the current enclosed by the loop. Then, Eq. (7.17) reduces to,

$$BL = \mu_0 I_e \qquad [7.17(b)]$$

ANDRE AMPERE (1775 –1836)

Andre Ampere (1775 – 1836) Andre Marie Ampere was a French physicist, mathematician and chemist who founded the science of electrodynamics. Ampere was a child prodigy who mastered advanced mathematics by the age of 12. Ampere grasped the significance of Oersted's discovery. He carried out a large series of experiments to explore the relationship between current electricity and magnetism. These investigations culminated in 1827 with the publication of the 'Mathematical Theory of Electrodynamic Phenomena Deduced Solely from Experiments'. He hypothesised that *all* magnetic phenomena are due to circulating electric currents. Ampere was humble and absent- minded. He once forgot an invitation to dine with the Emperor Napoleon. He died of pneumonia at the age of 61. His gravestone bears the epitaph: *Tandem Felix* (Happy at last).

When there is a system with a symmetry such as for a *straight infinite current-carrying wire* in Fig. 7.15, the Ampere's law enables an easy evaluation of the magnetic field, much the same way Gauss' law helps in determination of the electric field. This is exhibited in the Example 7.9 below. The boundary of the loop chosen is a circle and magnetic field is tangential to the circumference of the circle. The law gives, for the left hand side of Eq. [7.17 (b)], B. $2\pi r$. We find that the magnetic field at a distance r outside the wire is *tangential* and given by

$$B \times 2\pi r = \mu_0 I,$$

$$B = \mu_0 I/(2\pi r) \qquad (7.18)$$

The above result for the infinite wire is interesting from several points of view.

(i) It implies that the field at every point on a circle of radius r, (with the wire along the axis), is same in magnitude. In other words, the magnetic field possesses what is called a *cylindrical symmetry*. The field that normally can depend on three coordinates depends only on one: r. Whenever there is symmetry, the solutions simplify.

(ii) The field direction at any point on this circle is tangential to it. Thus, the lines of constant magnitude of magnetic field form concentric circles. Notice now, in Fig. 7.1(c), the iron filings form concentric circles. These lines called *magnetic field lines* form closed loops. This is unlike the electrostatic field lines which originate from positive charges and end at negative charges. The expression for the magnetic field of a straight wire provides a theoretical justification to Oersted's experiments.

(iii) Another interesting point to note is that even though the wire is infinite, the field due to it at a non-zero distance is *not* infinite. It tends to blow up only when we come very close to the wire. The field is directly proportional to the current and inversely proportional to the distance from the (infinitely long) current source.

(iv) There exists a simple rule to determine the direction of the magnetic field due to a long wire. This rule, called the *right-hand rule**, is :

Grasp the wire in your right hand with your extended thumb pointing in the direction of the current. Your fingers will curl around in the direction of the magnetic field.

Ampere's circuital law is not new in content from Biot-Savart law. Both relate the magnetic field and the current, and both express the same physical consequences of a steady electrical current. Ampere's law is to Biot-Savart law, what Gauss's law is to Coulomb's law. Both, Ampere's and Gauss's law relate a physical quantity on the periphery or boundary (magnetic or electric field) to another physical quantity, namely, the source, in the interior (current or charge). We also note that Ampere's circuital law holds for steady currents which

do not fluctuate with time. The following example will help us understand what is meant by the term *enclosed* current.

***** Note that there are *two distinct* right-hand rules: One which gives the direction of **B** on the axis of current-loop and the other which gives direction of **B** for a straight conducting wire. Fingers and thumb play different roles in the two.

It should be noted that while Ampere's circuital law holds for any loop, it may not always facilitate an evaluation of the magnetic field in every case. For example, for the case of the circular loop discussed in Section 7.6, it cannot be applied to extract the simple expression $B = \mu_0 I/2R$ [Eq. (7.16)] for the field at the centre of the loop. However, there exists a large number of situations of high symmetry where the law can be conveniently applied. We shall use it in the next section to calculate the magnetic field produced by two commonly used and very useful magnetic systems: the *solenoid* and the *toroid*.

7.8 THE SOLENOID AND THE TOROID

The solenoid and the toroid are two pieces of equipment which generate magnetic fields. The television uses the solenoid to generate magnetic fields needed. The synchrotron uses a combination of both to generate the high magnetic fields required. In both, solenoid and toroid, we come across a situation of high symmetry where Ampere's law can be conveniently applied.

7.8.1 The solenoid

We shall discuss a long solenoid. By long solenoid we mean that the solenoid's length is large compared to its radius. It consists of a long wire wound in the form of a helix where the neighbouring turns are closely spaced. So each turn can be regarded as a circular loop. The net magnetic field is the vector sum of the fields due to all the turns. Enamelled wires are used for winding so that turns are insulated from each other.

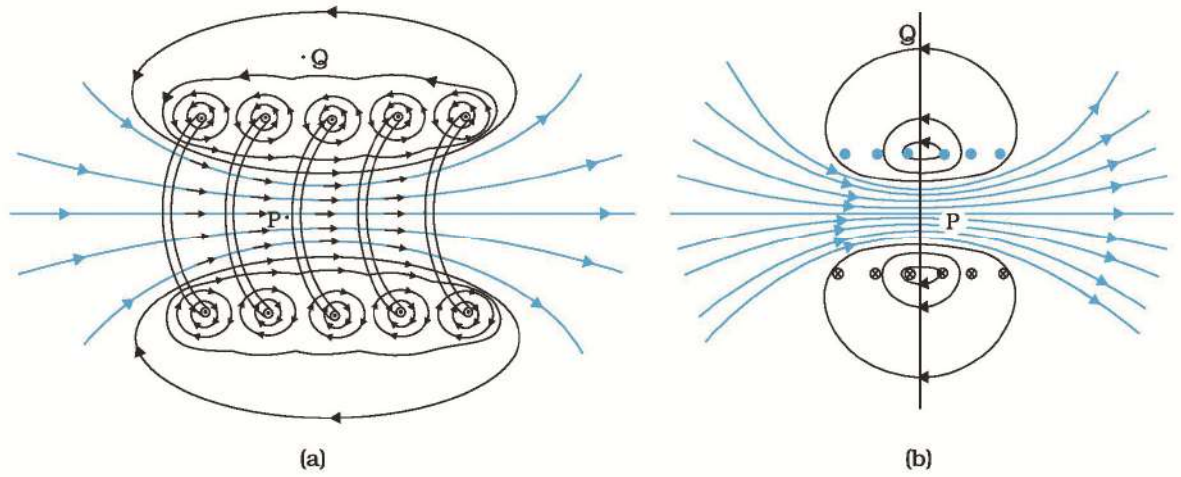


Fig. 7.17 (a) The magnetic field due to a section of the solenoid which has been stretched out for clarity. Only the exterior semi-circular part is shown. Notice how the circular loops between neighbouring turns tend to cancel. (b) The magnetic field of a finite solenoid.

Figure 7.17 displays the magnetic field lines for a finite solenoid. We show a section of this solenoid in an enlarged manner in Fig. 7.17(a). Figure 7.17(b) shows the entire finite solenoid with its magnetic field. In Fig. 7.17(a), it is clear from the circular loops that the field between two neighbouring turns vanishes. In Fig. 7.17(b), we see that the field at the interior mid-point P is uniform, strong and along the axis of the solenoid. The field at the exterior mid-point Q is weak and moreover is along the axis of the solenoid with no perpendicular or normal component. As the solenoid is made longer it appears like a long cylindrical metal sheet. Figure 4.18 represents this idealised picture. The field outside the

solenoid approaches zero. We shall assume that the field outside is zero. The field inside becomes everywhere parallel to the axis.

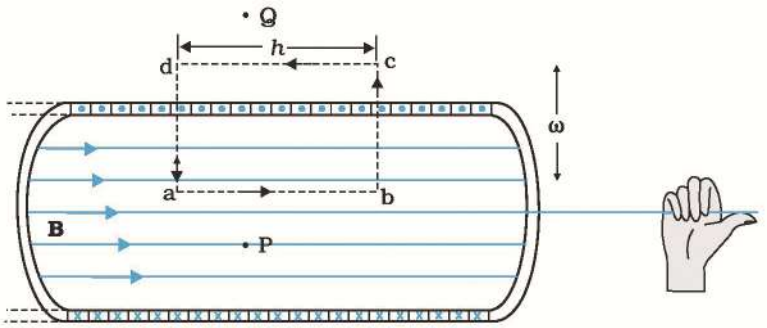


Fig. 7.18 The magnetic field of a very long solenoid. We consider a rectangular Amperian loop abcd to determine the field.

Consider a rectangular Amperian loop abcd. Along cd the field is zero as argued above. Along transverse sections bc and ad, the field component is zero. Thus, these two sections make no contribution. Let the field along ab be *B*. Thus, the relevant length of the Amperian loop is, $L = h$.

Let *n* be the number of turns per unit length, then the total number of turns is *nh*. The enclosed current is, $I_e = I\ (n\ h)$, where *I* is the current in the solenoid. From Ampere's circuital law [Eq. 7.17 (b)]

$$BL = \mu_0 I_e, \qquad B\ h = \mu_0 I\ (n\ h)$$

$$B = \mu_0 n\ I \tag{7.20}$$

The direction of the field is given by the right-hand rule. The solenoid is commonly used to obtain a uniform magnetic field. We shall see in the next chapter that a large field is possible by inserting a soft iron core inside the solenoid.

7.8.2 The toroid

The toroid is a hollow circular ring on which a large number of turns of a wire are closely wound. It can be viewed as a solenoid which has been bent into a circular shape to close on itself. It is shown in Fig. 7.19(a) carrying a current *I*. We shall see that the magnetic field in the open space inside (point P) and exterior to the toroid (point Q) is zero. The field **B** inside the toroid is constant in magnitude for the *ideal* toroid of closely wound turns.

Figure 7.19(b) shows a sectional view of the toroid. The direction of the magnetic field inside is clockwise as per the right-hand thumb rule for circular loops. Three circular Amperian loops 1, 2 and 3 are shown by dashed lines. By symmetry, the magnetic field should be tangential to each of them and constant in magnitude for a given loop. The circular areas bounded by loops 2 and 3 both cut the toroid: so that each turn of current carrying wire is cut once by the loop 2 and twice by the loop 3.

Let the magnetic field along loop 1 be B_1 in magnitude. Then in Ampere's circuital law [Eq. 7.17(a)], $L = 2\pi r_1$. However, the loop encloses no current, so $I_e = 0$. Thus,

$$B_1\ (2\ \pi r_1) = \mu_0(0), \qquad B_1 = 0$$

Thus, the magnetic field at any point P in the open space inside the toroid is zero.

We shall now show that magnetic field at Q is likewise zero. Let the magnetic field along loop 3 be B_3. Once again from Ampere's law $L = 2\ \pi r_3$. However, from the sectional cut, we see that the current coming out of the plane of the paper is cancelled exactly by the current going into it. Thus, $I_e = 0$, and $B_3 = 0$. Let the magnetic field inside the solenoid be *B*. We

shall now consider the magnetic field at S. Once again we employ Ampere's law in the form of Eq. [7.17 (a)]. We find, $L = 2\pi r$.

The current enclosed I_e is (for N turns of toroidal coil) $N I$.

$$B\,(2\pi r) = \mu_0 NI$$

$$B = \frac{\mu_0 NI}{2\pi r} \qquad (7.21)$$

We shall now compare the two results: for a toroid and solenoid. We re-express Eq. (7.21) to make the comparison easier with the solenoid result given in Eq. (7.20). Let r be the average radius of the toroid and n be the number of turns per unit length.

Then $N = 2\pi r\ n$ = *(average) perimeter of the toroid* × *number of turns per unit length* and thus,

$$B = \mu_0 n I, \qquad (7.22)$$

i.e., the result for the solenoid!

In an ideal toroid the coils are circular. In reality the turns of the toroidal coil form a helix and there is always a small magnetic field external to the toroid.

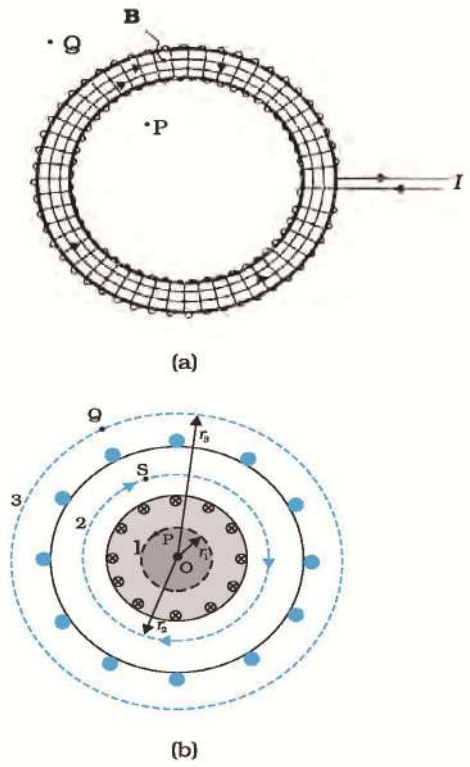
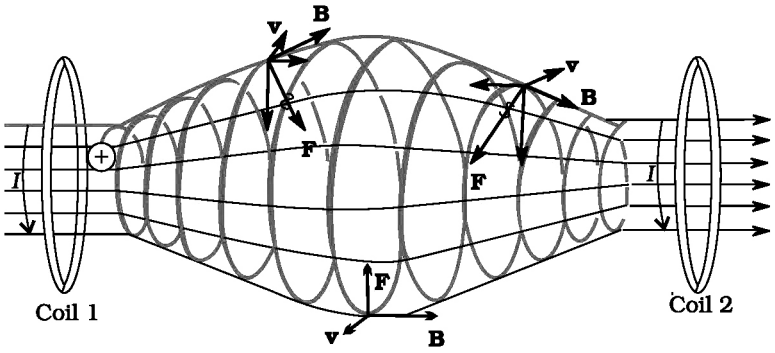


Fig.7.19 (a) A toroid carrying a current I. (b) A sectional view of the toroid. The magnetic field can be obtained at an arbitrary distance r from the centre O of the toroid by Ampere's circuital law. The dashed lines labelled 1, 2 and 3 are three circular Amperian loops.

MAGNETIC CONFINEMENT

We have seen in Section 7.3 (see also the box on helical motion of charged particles earlier in this chapter) that orbits of charged particles are helical. If the magnetic field is non-uniform, but does not change much during one circular orbit, then the radius of the helix will decrease as it enters stronger magnetic field and the radius will increase when it enters weaker magnetic fields. We consider two solenoids at a distance from each other, enclosed in an evacuated container (see figure below where we have not shown the container). Charged particles moving in the region between the two solenoids will start with a small radius. The radius will increase as field decreases and the radius will decrease again as field due to the second solenoid takes over. The solenoids act as a mirror or reflector. [See the direction of **F** as the particle approaches coil 2 in the figure. It has a horizontal component against the forward motion.] This makes the particles turn back when they approach the solenoid. Such an arrangement will act like *magnetic bottle* or magnetic container. The particles will never touch the sides of the container. Such magnetic bottles are of great use in confining the high energy plasma in fusion experiments. The plasma will destroy any other form of material container because of its high temperature. Another useful container is a toroid. Toroids are expected to play a key role in the *tokamak*, equipment for plasma confinement in fusion power reactors. There is an international collaboration called the *International Thermonuclear Experimental Reactor* (ITER), being set up in France, for achieving

controlled fusion, of which India is a collaborating nation. For details of ITER collaboration and the project, you may visit http://www.iter.org.



7.9 FORCE BETWEEN TWO PARALLEL CURRENTS, THE AMPERE

We have learnt that there exists a magnetic field due to a conductor carrying a current which obeys the Biot-Savart law. Further, we have learnt that an external magnetic field will exert a force on a current-carrying conductor. This follows from the Lorentz force formula. Thus, it is logical to expect that two current-carrying conductors placed near each other will exert (magnetic) forces on each other. In the period 1820-25, Ampere studied the nature of this magnetic force and its dependence on the magnitude of the current, on the shape and size of the conductors, as well as, the distances between the conductors. In this section, we shall take the simple example of two parallel current carrying conductors, which will perhaps help us to appreciate Ampere's painstaking work.

Figure 7.20 shows two long parallel conductors a and b separated by a distance *d* and carrying (parallel) currents *I*a and *I*b, respectively. The conductor 'a' produces, the same magnetic field **B**a at all points along the conductor 'b'. The right-hand rule tells us that the direction of this field is downwards (when the conductors are placed horizontally). Its magnitude is given by Eq. [7.19(a)] or from Ampere's circuital law,

$$B_a = \frac{\mu_0 I_a}{2\pi d}$$

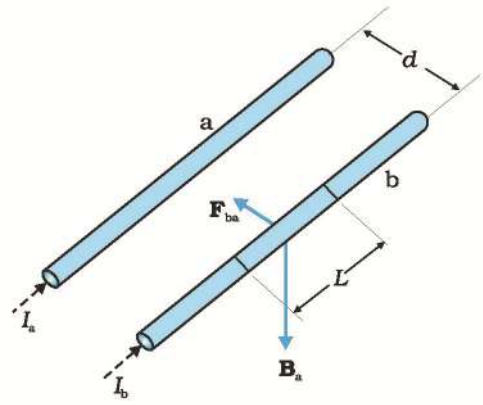


FIGURE 7.20 Two long straight parallel conductors carrying steady currents I_a and I_b and separated by a distance d. $\mathbf{B}_a$ is the magnetic field set up by conductor 'a' at conductor 'b'.

The conductor 'b' carrying a current I_b will experience a sideways force due to the field $\mathbf{B}_a$. The direction of this force is towards the conductor 'a' (Verify this). We label this force as $\mathbf{F}_{ba}$, the force on a segment *L* of 'b' due to 'a'. The magnitude of this force is given by Eq. (7.4),

$$F_{ba} = I_b L B_a$$

$$= \frac{\mu_0 I_a I_b}{2\pi d} L \qquad (7.23)$$

It is of course possible to compute the force on 'a' due to 'b'. From considerations similar to above we can find the force $\mathbf{F}_{ab}$, on a segment of length *L* of 'a' due to the current in 'b'. It is equal in magnitude to F_{ba}, and directed towards 'b'. Thus,

$$\mathbf{F}_{ba} = -\mathbf{F}_{ab} \quad (7.24)$$

Note that this is consistent with Newton's third Law. Thus, at least for parallel conductors and steady currents, we have shown that the Biot-Savart law and the Lorentz force yield results in accordance with Newton's third Law*****.

We have seen from above that currents flowing in the same direction attract each other. One can show that oppositely directed currents repel each other. Thus,

Parallel currents attract, and antiparallel currents repel.

This rule is the opposite of what we find in electrostatics. Like (same sign) charges repel each other, but like (parallel) currents attract each other.

Let f_{ba} represent the magnitude of the force $\mathbf{F}_{ba}$ per unit length. Then, from Eq. (7.23),

$$f_{ba} = \frac{\mu_0 I_a I_b}{2\pi d} \quad (7.25)$$

The above expression is used to define the ampere (A), which is one of the seven SI base units.

The *ampere* is the value of that steady current which, when maintained in each of the two very long, straight, parallel conductors of negligible cross-section, and placed one metre apart in vacuum, would produce on each of these conductors a force equal to 2×10^{-7} newtons per metre of length.

This definition of the ampere was adopted in 1946. It is a theoretical definition. In practice, one must eliminate the effect of the earth's magnetic field and substitute very long wires by multiturn coils of appropriate geometries. An instrument called the current balance is used to measure this mechanical force.

The SI unit of charge, namely, the coulomb, can now be defined in terms of the ampere.

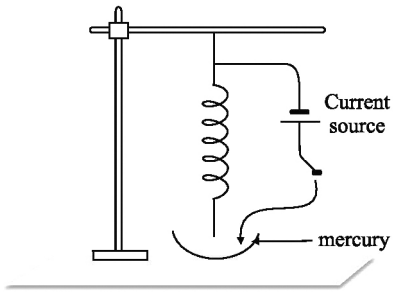
When a steady current of 1A is set up in a conductor, the quantity of charge that flows through its cross-section in 1s is one coulomb (1C).

***** It turns out that when we have time-dependent currents and/or charges in motion, Newton's third law may not hold for forces between charges and/or conductors. An essential consequence of the Newton's third law in mechanics is conservation of momentum of an isolated system. This, however, holds even for the case of time-dependent situations with electromagnetic fields, provided the momentum carried by fields is also taken into account.

ROGET'S SPIRAL FOR ATTRACTION BETWEEN PARALLEL CURRENTS

Magnetic effects are generally smaller than electric effects. As a consequence, the force between currents is rather small, because of the smallness of the factor μ. Hence it is difficult to demonstrate attraction or repulsion between currents. Thus, for 5 A current in each wire at a separation of 1cm, the force per metre would be 5×10^{-4} N, which is about 50 mg weight. It would be like pulling a wire by a string going over a pulley to which a 50 mg weight is attached. The displacement of the wire would be quite unnoticeable.

With the use of a soft spring, we can increase the effective length of the parallel current and by using mercury, we can make the displacement of even a few mm observable very dramatically. You will also need a constant-current supply giving a constant current of about 5 A. Take a soft spring whose natural period of oscillations is about 0.5 – 1s. Hang it vertically and attach a pointed tip to its lower end, as shown in the figure here. Take some mercury in a dish and adjust the spring such that the tip is just above the mercury surface.



Take the DC current source, connect one of its terminals to the upper end of the spring, and dip the other terminal in mercury. If the tip of the spring touches mercury, the circuit is completed through mercury.

Let the DC source be put off to begin with. Let the tip be adjusted so that it just touches the mercury surface. Switch on the constant current supply, and watch the fascinating outcome. The spring shrinks with a jerk, the tip comes out of mercury (just by a mm or so), the circuit is broken, the current stops, the spring relaxes and tries to come back to its original position, the tip again touches mercury establishing a current in the circuit, and the cycle continues with tick, tick, tick,... In the beginning, you may require some small adjustments to get a good effect.

Keep your face away from mercury vapour as it is poisonous. Do not inhale mercury vapour for long.

7.10 TORQUE ON CURRENT LOOP, MAGNETIC DIPOLE

7.10.1 Torque on a rectangular current loop in a uniform magnetic field

We now show that a rectangular loop carrying a steady current I and placed in a uniform magnetic field experiences a torque. It does not experience a net force. This behaviour is analogous to that of electric dipole in a uniform electric field (Section 4.12).

We first consider the simple case when the rectangular loop is placed such that the uniform magnetic field **B** is in the plane of the loop. This is illustrated in Fig. 7.21(a).

The field exerts no force on the two arms AD and BC of the loop. It is perpendicular to the arm AB of the loop and exerts a force $\mathbf{F}_1$ on it which is directed into the plane of the loop. Its magnitude is,

$$F_1 = I\, b\, B$$

Similarly, it exerts a force $\mathbf{F}_2$ on the arm CD and $\mathbf{F}_2$ is directed out of the plane of the paper.

$$F_2 = I\, b\, B = F_1$$

Thus, the *net force* on the loop *is zero*. There is a torque on the loop due to the pair of forces $\mathbf{F}_1$ and $\mathbf{F}_2$. Figure 7.21(b) shows a view of the loop from the AD end. It shows that the torque on the loop tends to rotate it anticlockwise. This torque is (in magnitude),

$$\begin{aligned} \tau &= F_1 \frac{a}{2} + F_2 \frac{a}{2} \\ &= IbB\frac{a}{2} + IbB\frac{a}{2} = I(ab)B \\ &= I\, A\, B \end{aligned} \qquad (7.26)$$

where $A = ab$ is the area of the rectangle.

We next consider the case when the plane of the loop, is not along the magnetic field, but makes an angle with it. We take the angle between the field and the normal to the coil to be angle θ(The previous case corresponds to $\theta = p/2$). Figure 7.22 illustrates this general case. The forces on the arms BC and DA are equal, opposite, and act along the axis of the coil, which connects the centres of mass of BC and DA. Being collinear along the axis they cancel each other, resulting in no net force or torque.

The forces on arms AB and CD are $\mathbf{F}_1$ and $\mathbf{F}_2$. They too are equal and opposite, with magnitude,

$$F_1 = F_2 = I\, b\, B$$

But they are not collinear! This results in a couple as before. The torque is, however, less than the earlier case when plane of loop was along the magnetic field. This is because the perpendicular distance between the forces of the couple has decreased. Figure 7.22(b) is a view of the arrangement from the AD end and it illustrates these two forces constituting a couple. The magnitude of the torque on the loop is,

$$t = F_1 \frac{a}{2} \sin\theta + F_2 \frac{a}{2} \sin\theta$$

$$= I\, ab\, B \sin\theta$$

$$= I\, A\, B \sin\theta \qquad (7.27)$$

As $\theta \to 0$, the perpendicular distance between the forces of the couple also approaches zero. This makes the forces collinear and the net force and torque zero. The torques in Eqs. (7.26) and (7.27) can be expressed as vector product of the magnetic moment of the coil and the magnetic field. We define the *magnetic moment* of the current loop as,

$$\mathbf{m} = I\, \mathbf{A} \qquad (7.28)$$

where the direction of the area vector **A** is given by the right-hand thumb rule and is directed into the plane of the paper in Fig. 7.21. Then as the angle between **m** and **B** is θ, Eqs. (7.26) and (7.27) can be expressed by one expression

$$t = \mathbf{m} \times \mathbf{B} \qquad (7.29)$$

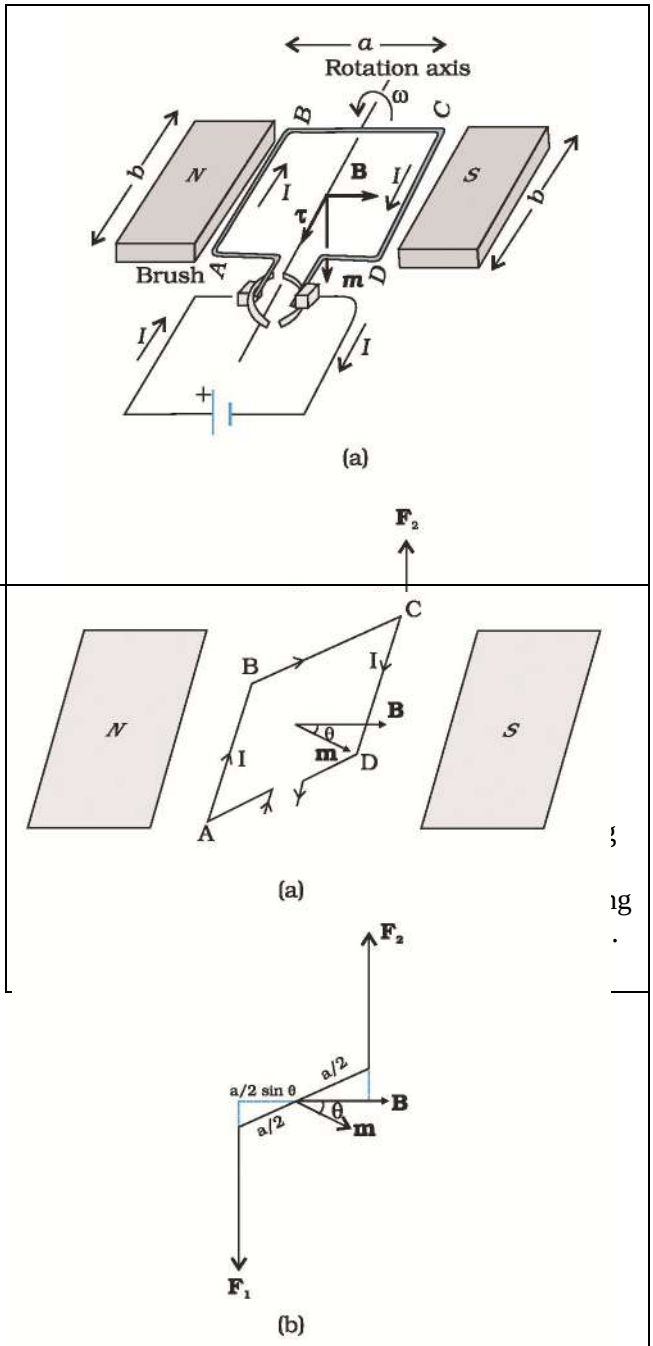


FIGURE 7.22 (a) The area vector of the loop ABCD makes an arbitrary angle θwith the magnetic field. (b) Top view of the loop. The forces $\mathbf{F}_1$ and $\mathbf{F}_2$ acting on the arms AB and CD are indicated.

This is analogous to the electrostatic case (Electric dipole of dipole moment **p**e in an electric field **E**).

$$\mathbf{t} = \mathbf{p}_e \times \mathbf{E}$$

As is clear from Eq. (7.28), the dimensions of the magnetic moment are $[A][L^2]$ and its unit is Am^2.

From Eq. (7.29), we see that the torque τvanishes when **m** is either parallel or antiparallel to the magnetic field **B**. This indicates a state of equilibrium as there is no torque on the coil (this also applies to any object with a magnetic moment **m**). When **m** and **B** are

parallel the equilibrium is a stable one. Any small rotation of the coil produces a torque which brings it back to its original position. When they are antiparallel, the equilibrium is unstable as any rotation produces a torque which increases with the amount of rotation. The presence of this torque is also the reason why a small magnet or any magnetic dipole aligns itself with the external magnetic field.

If the loop has N closely wound turns, the expression for torque, Eq. (7.29), still holds, with

$$\mathbf{m} = N I \mathbf{A} \qquad (7.30)$$

7.10.2 Circular current loop as a magnetic dipole

In this section, we shall consider the elementary magnetic element : the current loop. We shall show that the magnetic field (at large distances) due to current in a circular current loop is very similar in behaviour to the electric field of an electric dipole. In Section 7.6, we have evaluated the magnetic field on the axis of a circular loop, of a radius R, carrying a steady current I. The magnitude of this field is [(Eq. (7.15)],

$$B = \frac{\mu_0 I R^2}{2\left(x^2 + R^2\right)^{3/2}}$$

and its direction is along the axis and given by the right-hand thumb rule (Fig. 7.12). Here, x is the distance along the axis from the centre of the loop. For $x >> R$, we may drop the R^2 term in the denominator. Thus,

$$B = \frac{\mu_0 I R^2}{2x^3}$$

Note that the area of the loop $A = \pi R^2$. Thus,

$$B = \frac{\mu_0 I A}{2\pi x^3}$$

As earlier, we define the magnetic moment **m** to have a magnitude IA,

$\boldsymbol{m} = I\,\mathbf{A}$. Hence,

$$B = \frac{\mu_0 m}{2\pi x^3}$$

$$= \frac{\mu_0}{4\pi}\frac{2\mathbf{m}}{x^3} \qquad [7.31(a)]$$

The expression of Eq. [7.31(a)] is very similar to an expression obtained earlier for the electric field of a dipole. The similarity may be seen if we substitute,

$\mu_0 \rightarrow 1/e_0$

$\mathbf{m} \rightarrow \mathbf{p}$e (electrostatic dipole)

$\mathbf{B} \rightarrow \mathbf{E}$ (electrostatic field)

We then obtain,

$$\mathbf{E} = \frac{2\mathbf{p}_e}{4\pi\varepsilon_0 x^3}$$

which is precisely the field for an electric dipole at a point on its axis. considered in Chapter 1, Section 1.10 [Eq. (4.20)].

It can be shown that the above analogy can be carried further. We had found in Chapter 1 that the electric field on the perpendicular bisector of the dipole is given by [See Eq.(4.21)],

$$E = \frac{\mathbf{p}_e}{4\pi\varepsilon_0 x^3}$$

where x is the distance from the dipole. If we replace $\mathbf{p} \rightarrow \mathbf{m}$ and $\mu_0 \rightarrow 1/\varepsilon_0$ in the above expression, we obtain the result for **B** for a point *in the plane of the loop* at a distance x from the centre. For $x >> R$,

$$\mathbf{B} = \frac{\mu_0}{4\pi}\frac{\mathbf{m}}{x^3}; \qquad x >> R \qquad [7.31(b)]$$

The results given by Eqs. [7.31(a)] and [7.31(b)] become exact for a *point* magnetic dipole.

The results obtained above can be shown to apply to any planar loop: a planar current loop is equivalent to a magnetic dipole of dipole moment $\mathbf{m} = I\,\mathbf{A}$, which is the analogue of electric dipole moment **p**. Note, however, a fundamental difference: an electric dipole is built up of two elementary units — the charges (or electric monopoles). In magnetism, a magnetic dipole (or a current loop) is the most elementary element. The equivalent of electric charges, i.e., magnetic monopoles, are not known to exist.

We have shown that a current loop (i) produces a magnetic field (see Fig. 7.12) and behaves like a magnetic dipole at large distances, and (ii) is subject to torque like a magnetic needle. This led Ampere to suggest that all magnetism is due to circulating currents. This seems to be partly true and no magnetic monopoles have been seen so far. However, elementary particles such as an electron or a proton also carry an *intrinsic* magnetic moment, not accounted by circulating currents.

7.10.3 The magnetic dipole moment of a revolving electron

In Chapter 12 we shall read about the Bohr model of the hydrogen atom. You may perhaps have heard of this model which was proposed by the Danish physicist Niels Bohr in 1911 and was a stepping stone to a new kind of mechanics, namely, quantum mechanics. In the Bohr model, the electron (a negatively charged particle) revolves around a positively charged nucleus much as a planet revolves around the sun. The force in the former case is electrostatic (Coulomb force) while it is gravitational for the planet-Sun case. We show this Bohr picture of the electron in Fig. 7.23.

The electron of charge $(-e)$ ($e = +1.6 \times 10^{-19}$ C) performs uniform circular motion around a stationary heavy nucleus of charge $+Ze$. This constitutes a current I, where,

$$I = \frac{e}{T} \qquad (7.32)$$

and T is the time period of revolution. Let r be the orbital radius of the electron, and v the orbital speed. Then,

$$T = \frac{2\pi r}{v} \qquad (7.33)$$

Substituting in Eq. (7.32), we have $I = ev/2\pi r$.

There will be a magnetic moment, usually denoted by μ_l , associated with this circulating current. From Eq. (7.28) its magnitude is, $\mu_l = I\pi r^2 = evr/2$. The direction of this magnetic moment is into the plane of the paper in Fig. 7.23. [This follows from the right-hand rule discussed earlier and the fact that the negatively charged electron is moving anticlockwise, leading to a clockwise current].

Multiplying and dividing the right-hand side of the above expression by the electron mass m_e, we have,

$$\mu_l = \frac{e}{2m_e}(m_e v r)$$

$$= \frac{e}{2m_e} l \qquad [7.34(a)]$$

Here, l is the magnitude of the angular momentum of the electron about the central nucleus ("orbital" angular momentum). Vectorially,

$$\boldsymbol{\mu}_l = -\frac{e}{2m_e}\boldsymbol{l} \qquad [7.34(b)]$$

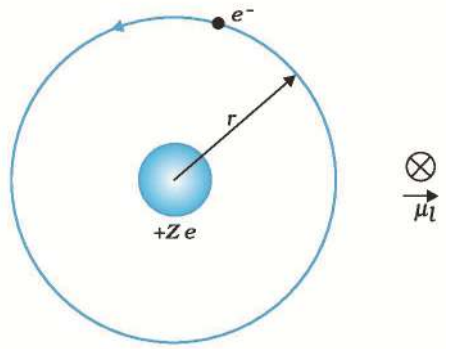


FIGURE 7.23 In the Bohr model of hydrogen-like atoms, the negatively charged electron is revolving with uniform speed around a centrally placed positively charged ($+Z\,e$) nucleus. The uniform circular motion of the electron constitutes a current. The direction of the magnetic moment is into the plane of the paper and is indicated separately by ⊗.

The negative sign indicates that the angular momentum of the electron is opposite in direction to the magnetic moment. Instead of electron with charge ($-e$), if we had taken a particle with charge ($+q$), the angular momentum and magnetic moment would be in the same direction. The ratio

$$\frac{\mu_l}{l} = \frac{e}{2m_e} \qquad (7.35)$$

is called the *gyromagnetic ratio* and is a constant. Its value is 8.8×10^{10} C /kg for an electron, which has been verified by experiments.

Conversion of galvanometer into ammeter and voltmeter:

The fact that even at an atomic level there is a magnetic moment, confirms Ampere's bold hypothesis of atomic magnetic moments. This according to Ampere, would help one to explain the magnetic properties of materials. Can one assign a value to this atomic dipole moment? The answer is Yes. One can do so within the Bohr model. Bohr hypothesised that the angular momentum assumes a discrete set of values, namely,

$$l = \frac{nh}{2\pi} \qquad (7.36)$$

where n is a natural number, n = 1, 2, 3, ... and h is a constant named after Max Planck (Planck's constant) with a value $h = 6.626 \times 10^{-34}$ J s. This condition of discreteness is called the *Bohr quantisation condition.* We shall discuss it in detail in Chapter 12. Our aim here is merely to use it to calculate the elementary dipole moment. Take the value n = 1, we have from Eq. (7.34) that,

$$(\mu_l)_{\min} = \frac{e}{4\pi m_e} h$$

$$= \frac{1.60\times10^{-19}\times6.63\times10^{-34}}{4\times3.14\times9.11\times10^{-31}}$$

$$= 9.27 \times 10^{-24}\ \text{Am}^2 \qquad (7.37)$$

where the subscript 'min' stands for minimum. This value is called the *Bohr magneton.*

Any charge in uniform circular motion would have an associated magnetic moment given by an expression similar to Eq. (7.34). This dipole moment is labelled as the *orbital magnetic moment*. Hence, the subscript '*l*' in *ml*. Besides the orbital moment, the electron has an *intrinsic* magnetic moment, which has the same numerical value as given in Eq. (7.37). It is called the *spin magnetic moment*. But we hasten to add that it is not as though the electron is spinning. The electron is an elementary particle and it does not have an axis to spin around like a top or our earth. Nevertheless, it does possess this *intrinsic* magnetic moment. The microscopic roots of magnetism in iron and other materials can be traced back to this intrinsic spin magnetic moment.

7.11 THE MOVING COIL GALVANOMETER

Currents and voltages in circuits have been discussed extensively in Chapters 3. But how do we measure them? How do we claim that current in a circuit is 1.5 A or the voltage drop across a resistor is 1.2 V? Figure 7.24 exhibits a very useful instrument for this purpose: the *moving coil galvanometer* (MCG). It is a device whose principle can be understood on the basis of our discussion in Section 7.10.

The galvanometer consists of a coil, with many turns, free to rotate about a fixed axis (Fig. 7.24), in a uniform radial magnetic field. There is a cylindrical soft iron core which not only makes the field radial but also increases the strength of the magnetic field. When a current flows through the coil, a torque acts on it. This torque is given by Eq. (7.26) to be

$$\tau = NI\,AB$$

where the symbols have their usual meaning. Since the field is radial by design, we have taken $\sin\theta = 1$ in the above expression for the torque. The magnetic torque *NIAB* tends to rotate the coil. A spring Sp provides a counter torque $k\phi$ that balances the magnetic torque *NIAB;* resulting in a steady angular deflection ϕ. In equilibrium

$$k\phi = NI\,AB$$

where *k* is the torsional constant of the spring; i.e. the restoring torque per unit twist. The deflection ϕ is indicated on the scale by a pointer attached to the spring. We have

$$\phi = \left(\frac{NAB}{k}\right) I \qquad (7.38)$$

The quantity in brackets is a constant for a given galvanometer.

The galvanometer can be used in a number of ways. It can be used as a detector to check if a current is flowing in the circuit. We have come across this usage in the Wheatstone's bridge arrangement. In this usage the neutral position of the pointer (when no current is flowing through the galvanometer) is in the middle of the scale and not at the left end as shown in Fig.7.24. Depending on the direction of the current, the pointer's deflection is either to the right or the left.

The galvanometer cannot as such be used as an ammeter to measure the value of the current in a given circuit. This is for two reasons: (i) Galvanometer is a very sensitive device, it gives a full-scale deflection for a current of the order of *m*A. (ii) For measuring currents, the galvanometer has to be connected in series, and as it has a large resistance, this will change the value of the current in the circuit.

To overcome these difficulties, one attaches a small resistance r_s, called *shunt resistance*, in parallel with the galvanometer coil; so that most of the current passes through the shunt. The resistance of this arrangement is,

$$R_G\ r_s / (R_G + r_s) = r_s \text{ if } \qquad R_G >> r_s$$

If r_s has small value, in relation to the resistance of the rest of the circuit R_c, the effect of introducing the measuring instrument is also small and negligible. This arrangement is schematically shown in Fig. 7.25. The scale of this ammeter is calibrated and then graduated to read off the current value with ease. We define the *current sensitivity of the galvanometer as the deflection per unit current*. From Eq. (7.38) this current sensitivity is,

$$\frac{\phi}{I} = \frac{NAB}{k} \qquad (7.39)$$

A convenient way for the manufacturer to increase the sensitivity is to increase the number of turns *N*. We choose galvanometers having sensitivities of value, required by our experiment.

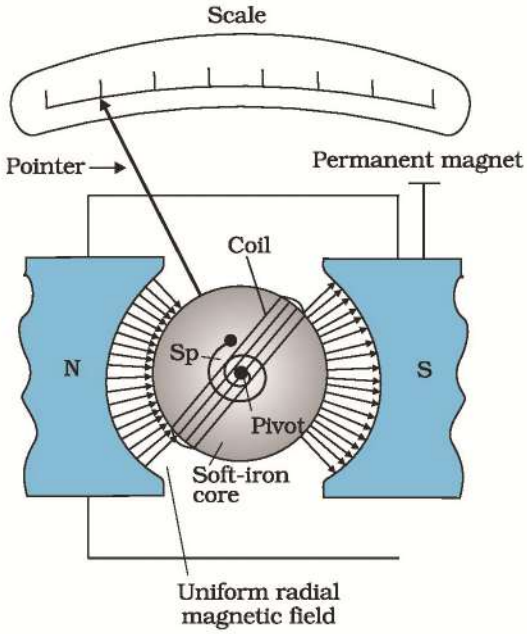


Fig.7.24 The moving coil galvanometer. Its elements are described in the text. Depending on the requirement, this device can be used as a current detector or for measuring the value of the current (ammeter) or voltage (voltmeter).

The galvanometer can also be used as a voltmeter to measure the voltage across a given section of the circuit. For this it must be connected *in parallel* with that section of the circuit. Further, it must draw a very small current, otherwise the voltage measurement will disturb the original set up by an amount which is very large. Usually we like to keep the disturbance due to the measuring device below one per cent. To ensure this, a large resistance *R* is connected *in series* with the galvanometer. This arrangement is schematically depicted in Fig.7.26. Note that the resistance of the voltmeter is now,

$$R_G + R = R : \text{large}$$

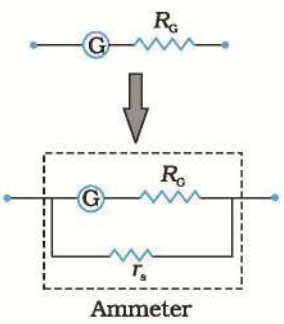


Fig. 7.25 Conversion of a galvanometer (G) to an ammeter by the introduction of a shunt resistance *rs* of very small value in parallel.

The scale of the voltmeter is calibrated to read off the voltage value with ease. We define the *voltage sensitivity as the deflection per unit voltage*. From Eq. (7.38),

$$\frac{\phi}{V} = \left(\frac{NAB}{k}\right)\frac{I}{V} = \left(\frac{NAB}{k}\right)\frac{I}{R} \qquad (7.40)$$

An interesting point to note is that increasing the current sensitivity may not necessarily increase the voltage sensitivity. Let us take Eq. (7.39) which provides a measure of current sensitivity. If $N \to 2N$, i.e., we double the number of turns, then

$$\frac{\phi}{I} \to 2\frac{\phi}{I}$$

Thus, the current sensitivity doubles. However, the resistance of the galvanometer is also likely to double, since it is proportional to the length of the wire. In Eq. (7.40), $N \rightarrow 2N$, and $R \rightarrow 2R$, thus the voltage sensitivity,

$$\frac{\phi}{V} \rightarrow \frac{\phi}{V}$$

remains unchanged. So in general, the modification needed for conversion of a galvanometer to an ammeter will be different from what is needed for converting it into a voltmeter.

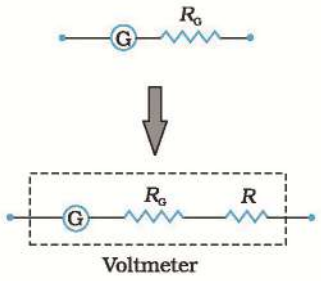


Fig. 7.26 Conversion of a galvanometer (G) to a voltmeter by the introduction of a resistance *R* of large value in series.

SUMMARY

1. The total force on a charge *q* moving with velocity **v** in the presence of magnetic and electric fields **B** and **E,** respectively is called the *Lorentz force*. It is given by the expression:

   $\mathbf{F} = q\,(\mathbf{v} \times \mathbf{B} + \mathbf{E})$

   The magnetic force $q\,(\mathbf{v} \times \mathbf{B})$ is normal to **v** and work done by it is zero.

2. A straight conductor of length *l* and carrying a steady current *I* experiences a force **F** in a uniform external magnetic field **B**,

   $\mathbf{F} = I\,\mathbf{l} \times \mathbf{B}$

   where $|\mathbf{l}| = l$ and the direction of **l** is given by the direction of the current.

3. In a uniform magnetic field **B**, a charge *q* executes a circular orbit in a plane normal to **B**. Its frequency of uniform circular motion is called the *cyclotron frequency* and is given by:

   $$\nu_c = \frac{qB}{2\pi m}$$

   This frequency is independent of the particle's speed and radius. This fact is exploited in a machine, the cyclotron, which is used to accelerate charged particles.

4. The *Biot-Savart* law asserts that the magnetic field d**B** due to an element d**l** carrying a steady current *I* at a point P at a distance *r* from the current element is:

   $$d\mathbf{B} = \frac{\mu_0}{4\pi} I \frac{d\mathbf{l} \times r}{r^3}$$

   To obtain the total field at P, we must integrate this vector expression over the entire length of the conductor.

5. The magnitude of the magnetic field due to a circular coil of radius *R* carrying a current *I* at an axial distance *x* from the centre is

   $$B = \frac{\mu_0 I R^2}{2\left(x^2 + R^2\right)^{3/2}}$$

   At the centre this reduces to

   $$B = \frac{\mu_0 I}{2R}$$

6. *Ampere's Circuital Law:* Let an open surface S be bounded by a loop C. Then the Ampere's law states that $\oint_C \mathbf{B}.d\mathbf{l} = \mu_0 I$ where *I* refers to the current passing through S.

   The sign of *I* is determined from the right-hand rule. We have discussed a simplified form of this law. If **B** is directed along the tangent to every point on the perimeter *L* of a closed curve and is constant in magnitude along perimeter then,

   $BL = \mu_0 I_e$

   where *Ie* is the net current enclosed by the closed circuit.

7. The magnitude of the magnetic field at a distance *R* from a long, straight wire carrying a current *I* is given by:

   $$B = \frac{\mu_0 I}{2\pi R}$$

   The field lines are circles concentric with the wire.

8. The magnitude of the field *B* inside a *long solenoid* carrying a current *I* is

   $B = \mu_0 nI$

   where *n* is the number of turns per unit length. For a *toroid* one obtains,

   $$B = \frac{\mu_0 NI}{2\pi r}$$

   where *N* is the total number of turns and *r* is the average radius.

9. Parallel currents attract and anti-parallel currents repel.

10. A planar loop carrying a current *I*, having *N* closely wound turns, and an area *A* possesses a magnetic moment **m** where,

    $\mathbf{m} = \mathrm{N}\, I\, \mathbf{A}$

    and the direction of **m** is given by the right-hand thumb rule : curl the palm of your right hand along the loop with the fingers pointing in the direction of the current. The thumb sticking out gives the direction of **m** (and **A**)

    When this loop is placed in a uniform magnetic field **B**, the force **F** on it is: *F = 0*

    And the torque on it is,

    $\mathrm{t} = \mathbf{m} \times \mathbf{B}$

    In a moving coil galvanometer, this torque is balanced by a counter-torque due to a spring, yielding

    $k\phi = NI\,AB$

    where $\phi$is the equilibrium deflection and *k* the torsion constant of the spring.

11. An electron moving around the central nucleus has a magnetic moment $\mu_l$given by:

    $$\mu_l = \frac{e}{2m} l$$

    where *l* is the magnitude of the angular momentum of the circulating electron about the central nucleus. The smallest value of μ_l is called the Bohr magneton μ_B and it is μ_B = 9.27 × 10 J/T

12. A moving coil galvanometer can be converted into a ammeter by introducing a shunt resistance r_s, of small value in parallel. It can be converted into a voltmeter by introducing a resistance of a large value in series.

Physical Quantity	Symbol	Nature	Dimensions	Units	Remarks
Permeability of free space	μ_0	Scalar	$[MLT^{-2} A]$	T m A^{-1}	$4\pi \times 10$ T m A^{-1}
Magnetic Field	**B**	Vector	$[M T^{-2} A^{-1}]$	T (telsa)	
Magnetic Moment	**m**	Vector	$[L^2 A]$	A m^2 or J/T	
Torsion Constant	*k*	Scalar	$[M L^2 T^{-2}]$	N m rad^{-1}	Appearsin MCG

POINTS TO PONDER

1. Electrostatic field lines originate at a positive charge and terminate at a negative charge or fade at infinity. Magnetic field lines always form closed loops.
2. The discussion in this Chapter holds only for steady currents which do not vary with time.

   When currents vary with time Newton's third law is valid only if momentum carried by the electromagnetic field is taken into account.
3. Recall the expression for the Lorentz force,

   $\mathbf{F} = q (\mathbf{v} \times \mathbf{B} + \mathbf{E})$

   This velocity dependent force has occupied the attention of some of the greatest scientific thinkers. If one switches to a frame with instantaneous velocity **v**, the magnetic part of the force vanishes. The motion of the charged particle is then explained by arguing that there exists an appropriate electric field in the new frame. We shall not discuss the details of this mechanism. However, we stress that the resolution of this paradox implies that electricity and magnetism are linked phenomena (*electromagnetism*) and that the Lorentz force expression *does not* imply a universal preferred frame of reference in nature.
4. Ampere's Circuital law is not independent of the Biot-Savart law. It can be derived from the Biot-Savart law. Its relationship to the Biot-Savart law is similar to the relationship between Gauss's law and Coulomb's law.

VERY SHORT ANSWER QUESTIONS (2 MARKS)

1. What is the importance of Oersted's experiment?
2. State Ampere's law and Biot-Savart law.
3. A circuit coil of radius 'r' having N turns carries a current "i". What is its magnetic moment?
4. What is the force on a conductor of length L carrying a current "i" placed in a magnetic field of induction B? When does it become maximum?
5. What is the force on a charged particle of charge "q" moving with a velocity "v" in a uniform magnetic field of induction B? When does it become maximum?
6. Distinguish between ammeter and voltmeter.
7. What is the principle of a moving coil galvanometer?
8. What is the smallest value of current that can be measured with a moving coil galvanometer?
9. How do you convert a moving coil galvanometer into an ammeter?
10. How do you convert a moving coil galvanometer into a voltmeter?

SHORT ANSWER QUESTIONS (4 MARKS)

1. State and explain Biot – Savart law.
2. State and explain Ampere's law.
3. Derive an expression for the magnetic induction at the centre of a current carrying circuit coil using Biot-Savart law.
4. Derive an expression for the magnetic induction at a point on the axis of of a current carrying circuit coil using Biot-Savart law.
5. Obtain an expression for the magnetic dipole moment of a current loop.
6. Derive an expression for the magnetic dipole moment of a revolving electron.
7. Explain how crossed **E** and **B** field serve as a velocity selector.
8. What are the basic components of a cyclotron? Mention its uses?

LONG ANSWER QUESTIONS (8 MARKS)

1. Deduce an expression for the force on a current carrying conductor placed in a magnetic field. Derive an expression for the force per unit length between two parallel current-carrying conductors.
2. Obtain an expression for the torque on a current carrying loop placed in a uniform magnetic field. Describe the construction and working of a moving coil galvanometer.
3. How can a galvanometer be converted to an ammeter? Why is the parallel resistance smaller that the galvanometer resistance?
4. How can a galvanometer be converted to a voltmeter? Why is the series resistance greater that the galvanometer resistance?
5. Derive an expression for the force acting between two very long parallel current-carrying conductors and hence define the Ampere.

CHAPTER 8

Magnetism and Matter

8.1 INTRODUCTION

Magnetic phenomena are universal in nature. Vast, distant galaxies, the tiny invisible atoms, humans and beasts all are permeated through and through with a host of magnetic fields from a variety of sources. The earth's magnetism predates human evolution. The word magnet is derived from the name of an island in Greece called *magnesia* where magnetic ore deposits were found, as early as 600 BC. Shepherds on this island complained that their wooden shoes (which had nails) at times stayed struck to the ground. Their iron-tipped rods were similarly affected. This attractive property of magnets made it difficult for them to move around. The directional property of magnets was also known since ancient times. A thin long piece of a magnet, when suspended freely, pointed in the north-south direction. A similar effect was observed when it was placed on a piece of cork which was then allowed to float in still water. The name *lodestone* (or *loadstone*) given to a naturally occurring ore of iron- magnetite means leading stone. The technological exploitation of this property is generally credited to the Chinese. Chinese texts dating 400 BC mention the use of magnetic needles for navigation on ships. Caravans crossing the Gobi desert also employed magnetic needles.

A Chinese legend narrates the tale of the victory of the emperor Huang-ti about four thousand years ago, which he owed to his craftsmen (whom nowadays you would call engineers). These 'engineers' built a chariot on which they placed a magnetic figure with arms outstretched. Figure 8.1 is an artist's description of this chariot. The figure swiveled around so that the finger of the statuette on it always pointed south. With this chariot, Huang-ti's troops were able to attack the enemy from the rear in thick fog, and to defeat them.

In the previous chapter we have learned that moving charges or electric currents produce magnetic fields. This discovery, which was made in the early part of the nineteenth century is credited to Oersted, Ampere, Biot and Savart, among others.

In the present chapter, we take a look at magnetism as a subject in its own right.

Some of the commonly known ideas regarding magnetism are:

(i) The earth behaves as a magnet with the magnetic field pointing approximately from the geographic south to the north.

(ii) When a bar magnet is freely suspended, it points in the north-south direction. The tip which points to the geographic north is called the *north pole* and the tip which points to the geographic south is called the *south pole* of the magnet.

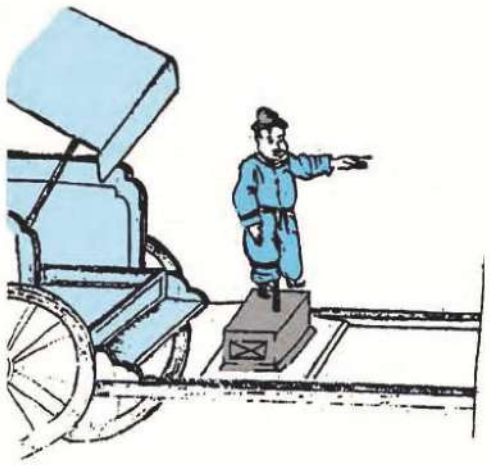


Fig. 8.1 The arm of the statuette mounted on the chariot always points south. This is an artist's sketch of one of the earliest known compasses, thousands of years old.

(iii) There is a repulsive force when north poles (or south poles) of two magnets are brought close together. Conversely, there is an attractive force between the north pole of one magnet and the south pole of the other.

(iv) We cannot isolate the north, or south pole of a magnet. If a bar magnet is broken into two halves, we get two similar bar magnets with somewhat weaker properties. Unlike electric charges, isolated magnetic north and south poles known as *magnetic monopoles* do not exist.

(v) It is possible to make magnets out of iron and its alloys.

We begin with a description of a bar magnet and its behaviour in an external magnetic field. We describe Gauss's law of magnetism. We then follow it up with an account of the earth's magnetic field. We next describe how materials can be classified on the basis of their magnetic properties. We describe para-, dia-, and ferromagnetism. We conclude with a section on electromagnets and permanent magnets.

8.2 THE BAR MAGNET

One of the earliest childhood memories of the famous physicist Albert Einstein was that of a magnet gifted to him by a relative. Einstein was fascinated, and played endlessly with it. He wondered how the magnet could affect objects such as nails or pins placed away from it and not in any way *connected* to it by a spring or string.

We begin our study by examining iron filings sprinkled on a sheet of glass placed over a short bar magnet. The arrangement of iron filings is shown in Fig. 8.2.

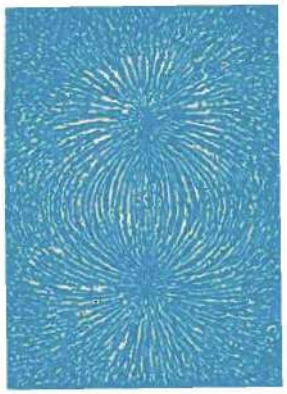


Fig. 8.2 The arrangement of iron filings surrounding a bar magnet. The pattern mimics magnetic field lines. The pattern suggests that the bar magnet is a magnetic dipole.

The pattern of iron filings suggests that the magnet has two poles similar to the positive and negative charge of an electric dipole. As mentioned in the introductory section, one pole is designated the *North pole* and the other, the *South pole*. When suspended freely, these poles point approximately towards the geographic north and south poles, respectively. A similar pattern of iron filings is observed around a current carrying solenoid.

8.2.1 The magnetic field lines

The pattern of iron filings permits us to plot the magnetic field lines*****. This is shown both for the bar-magnet and the current-carrying solenoid in Fig. 8.3. For comparison refer to the Chapter 4, Figure 4.17(d). Electric field lines of an electric dipole are also displayed in Fig. 8.3(c). The magnetic field lines are a visual and intuitive realisation of the magnetic field. Their properties are:

(i) The magnetic field lines of a magnet (or a solenoid) form continuous closed loops. This is unlike the electric dipole where these field lines begin from a positive charge and end on the negative charge or escape to infinity.

(ii) The tangent to the field line at a given point represents the direction of the net magnetic field **B** at that point.

***** In some textbooks the magnetic field lines are called *magnetic lines of force*. This nomenclature is avoided since it can be confusing. Unlike electrostatics the field lines in magnetism do not indicate the direction of the force on a (moving) charge.

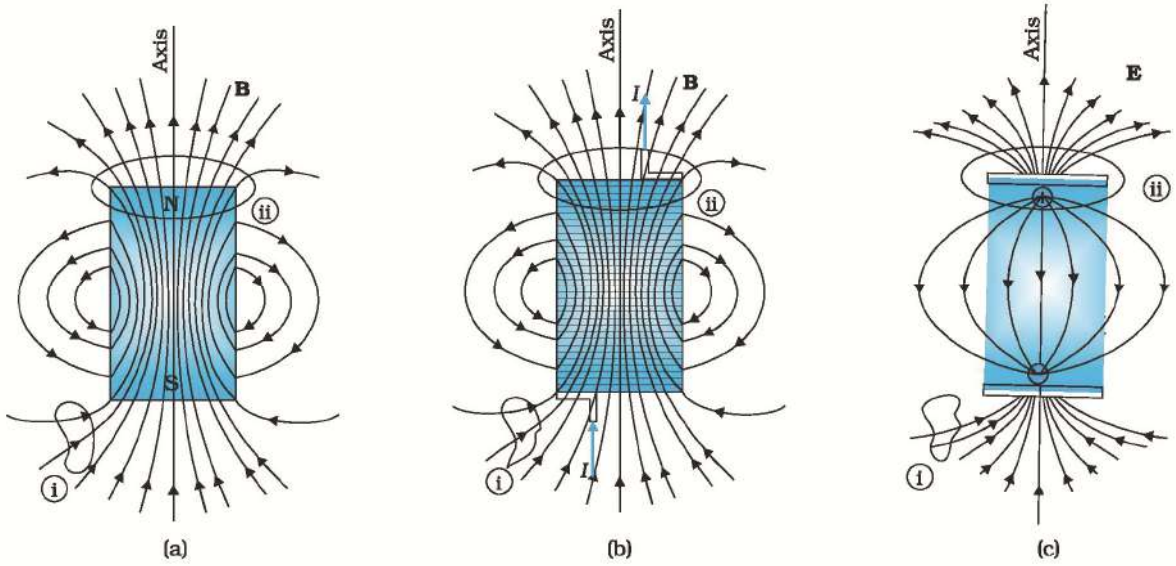


Fig. 8.3 The field lines of (a) a bar magnet, (b) a current-carrying finite solenoid and (c) electric dipole. At large distances, the field lines are very similar. The curves labelled i and ii are closed Gaussian surfaces.

(iii) The larger the number of field lines crossing per unit area, the stronger is the magnitude of the magnetic field **B**. In Fig. 8.3(a), **B** is larger around region ii than in region i .

(iv) The magnetic field lines do not intersect, for if they did, the direction of the magnetic field would not be unique at the point of intersection.

One can plot the magnetic field lines in a variety of ways. One way is to place a small magnetic compass needle at various positions and note its orientation. This gives us an idea of the magnetic field direction at various points in space.

8.2.2 Bar magnet as an equivalent solenoid

In the previous chapter, we have explained how a current loop acts as a magnetic dipole (Section 7.10). We mentioned Ampere's hypothesis that all magnetic phenomena can be explained in terms of circulating currents. Recall that the magnetic dipole moment **m** associated with a current loop was defined to be $\mathbf{m} = NI\,\mathbf{A}$ where N is the number of turns in the loop, I the current and **A** the area vector (Eq. 7.30).

The resemblance of magnetic field lines for a bar magnet and a solenoid suggest that a bar magnet may be thought of as a large number of circulating currents in analogy with a solenoid. Cutting a bar magnet in half is like cutting a solenoid. We get two smaller solenoids with weaker magnetic properties. The field lines remain continuous, emerging from one face of the solenoid and entering into the other face. One can test this analogy by moving a small compass needle in the neighbourhood of a bar magnet and a current-carrying finite solenoid and noting that the deflections of the needle are similar in both cases.

To make this analogy more firm we calculate the axial field of a finite solenoid depicted in Fig. 8.4 (a). We shall demonstrate that at large distances this axial field resembles that of a bar magnet.

Let the solenoid of Fig. 8.4(a) consists of n turns per unit length. Let its length be $2l$ and radius a. We can evaluate the axial field at a point P, at a distance r from the centre O of the solenoid. To do this, consider a circular element of thickness dx of the solenoid at a distance x from its centre. It consists of $n\,dx$ turns. Let I be the current in the solenoid. In Section 4.6 of the previous chapter we have calculated the magnetic field on the axis of a circular current loop. From Eq. (7.13), the magnitude of the field at point P due to the circular element is

$$dB = \frac{\mu_0 n dx I a^2}{2\left[(r-x)^2 + a^2\right]^{3/2}}$$

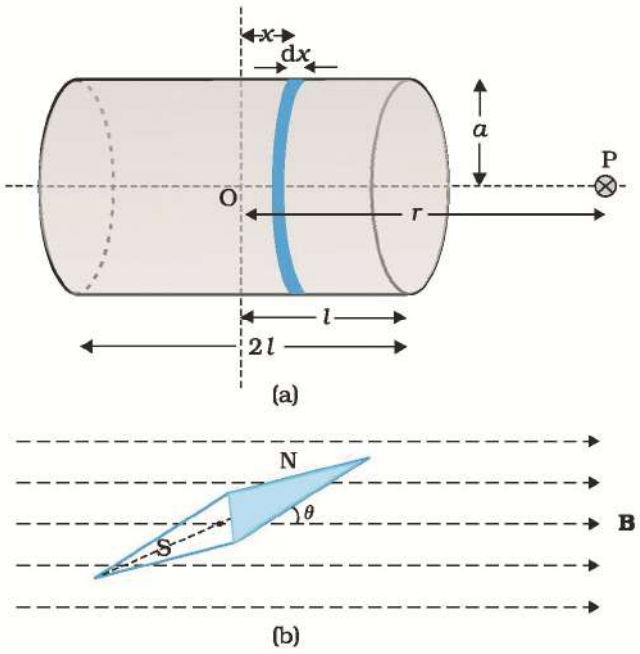


Fig. 8.4 Calculation of (a) The axial field of a finite solenoid in order to demonstrate its similarity to that of a bar magnet. (b) A magnetic needle in a uniform magnetic field **B**. The arrangement may be used to determine either **B** or the magnetic moment **m** of the needle.

The magnitude of the total field is obtained by summing over all the elements — in other words by integrating from $x = -l$ to $x = +l$. Thus,

$$B = \frac{\mu_0 n I a^2}{2}\int_{-l}^{l} \frac{dx}{\left[(r-x)^2 + a^2\right]^{3/2}}$$

This integration can be done by trigonometric substitutions. This exercise, however, is not necessary for our purpose. Note that the range of x is from $-l$ to $+l$. Consider the far axial field of the solenoid, i.e., $r >> a$ and $r >> l$. Then the denominator is approximated by

$$[(r-x)^2 + a^2]^{3/2} \approx r^3$$

$$\text{and } B = \frac{\mu_0 n I a^2}{2r^3}\int_{-l}^{l} dx$$

$$= \frac{\mu_0 n I}{2}\frac{2la^2}{r^3} \qquad (8.1)$$

Note that the magnitude of the magnetic moment of the solenoid is, $m = n\,(2\,l)\,I\,(\pi a^2)$ — (total number of turns × current × cross-sectional area). Thus,

$$B = \frac{\mu_0}{4\pi}\frac{2m}{r^3} \qquad (8.2)$$

This is also the far axial magnetic field of a bar magnet which one may obtain experimentally. Thus, a bar magnet and a solenoid produce similar magnetic fields. The magnetic moment of a bar magnet is thus equal to the magnetic moment of an equivalent solenoid that produces the same magnetic field.

Some textbooks assign a *magnetic charge* (also called *pole strength*) $+q_m$ to the north pole and $-q_m$ to the south pole of a bar magnet of length $2l$, and magnetic moment $qm(2l)$. The field strength due to q_m at a distance r from it is given by $\frac{\mu_0 q_m}{4\pi r^2}$. The magnetic field due to the bar magnet 0 *m* is then obtained, both for the axial and the equatorial case, in a manner analogous to that of an electric dipole (Chapter 4). The method is simple and appealing. However, *magnetic monopoles do not exist*, and *we have avoided this approach for that reason*.

8.2.3 The dipole in a uniform magnetic field

The pattern of iron filings, i.e., the magnetic field lines gives us an approximate idea of the magnetic field **B**. We may at times be required to determine the magnitude of **B** accurately. This is done by placing a small compass needle of known magnetic moment **m** and moment of inertia**I** and allowing it to oscillate in the magnetic field. This arrangement is shown in Fig. 8.4(b).

The torque on the needle is [see Eq. (7.29)],

$$\tau = \mathbf{m} \times \mathbf{B} \tag{8.3}$$

In magnitude $\tau = mB \sin\theta$

Here $\tau$is restoring torque and θ is the angle between **m** and **B**.

Therefore, in equilibrium $\mathbf{I}\frac{d^2\theta}{dt^2} = -mB \sin\theta$

Negative sign with $mB \sin\theta$implies that restoring torque is in opposition to deflecting torque. For small values of θin radians, we approximate $\sin\theta \approx \theta$and get

$$\mathbf{I}\frac{d^2\theta}{dt^2} = -mB\,\theta$$

or, $$\frac{d^2\theta}{dt^2} = -\frac{mB}{I}\sin\theta$$

This represents a simple harmonic motion. The square of the angular frequency is $\omega^2 = mB/I$ and the time period is,

$$T = 2\pi\sqrt{\frac{I}{mB}} \tag{8.4}$$

or $$B = \frac{4\pi^2 I}{mT^2} \tag{8.5}$$

An expression for magnetic potential energy can also be obtained on lines similar to electrostatic potential energy.

The magnetic potential energy $U_m$is given by

$$U_m = \int \tau(\theta) d\theta$$

$$= \int mB \sin\theta \, d\theta = -mB\cos\theta$$

$$= -\mathbf{m}.\mathbf{B} \tag{8.6}$$

We have emphasised in Chapter 5 that the zero of potential energy can be fixed at one's convenience. Taking the constant of integration to be zero means fixing the zero of potential energy at $\theta = 90°$, i.e., when the needle is perpendicular to the field. Equation (8.6) shows that potential energy is minimum ($= -mB$) at $\theta = 0°$ (most stable position) and maximum ($= +mB$) at $\theta = 180°$ (most unstable position).

8.2.4 The electrostatic analog

Comparison of Eqs. (8.2), (8.3) and (8.6) with the corresponding equations for electric dipole (Chapter 4), suggests that magnetic field at large distances due to a bar magnet of magnetic moment **m** can be obtained from the equation for electric field due to an electric dipole of dipole moment **p**, by making the following replacements :

$$\mathbf{E} \rightarrow \mathbf{B}\ ,\ \mathbf{p} \rightarrow \mathbf{m}\ ,\ \frac{1}{4\pi\varepsilon_0} \rightarrow \frac{\mu_0}{4\pi}$$

In particular, we can write down the equatorial field ($\mathbf{B}_E$) of a bar magnet at a distance r, for $r >> l$, where l is the size of the magnet:

$$\mathbf{B}_E = -\frac{\mu_0 m}{4\pi r^3} \qquad (8.7)$$

Likewise, the axial field ($\mathbf{B}_A$) of a bar magnet for $r >> l$ is:

$$\mathbf{B}_A = \frac{\mu_0}{4\pi}\frac{2m}{r^3} \qquad (8.8)$$

Equation (8.8) is just Eq. (8.2) in the vector form. Table 8.1 summarises the analogy between electric and magnetic dipoles.

TABLE 8.1 THE DIPOLE ANALOGY

	Electrostatics	**Magnetism**
	$1/\varepsilon_0$	μ_0
Dipole moment	**p**	**m**
Equatorial Field for a short dipole	$-\mathbf{p}/4\pi\varepsilon_0 r^3$	$-\mu_0\mathbf{m}/4\pi r^3$
Axial Field for a short dipole	$2\mathbf{p}/4\pi\varepsilon_0 r^3$	$\mu_0 2\mathbf{m}/4\pi r^3$
External Field: torque	$\mathbf{p} \times \mathbf{E}$	$\mathbf{m} \times \mathbf{B}$
External Field: Energy	$-\mathbf{p}.\mathbf{E}$	$-\mathbf{m}.\mathbf{B}$

8.3 MAGNETISM AND GAUSS'S LAW

In Chapter 1, we studied Gauss's law for electrostatics. In Fig 8.3(c), we see that for a closed surface represented by i , the number of lines leaving the surface is equal to the number of lines entering it. This is consistent with the fact that no net charge is enclosed by the surface. However, in the same figure, for the closed surface ii , thereis a net outward flux, since it does include a net (positive) charge.

KARL FRIEDRICH GAUSS (1777 – 1855)

Karl Friedrich Gauss (1777 – 1855) He was a child prodigy and was gifted in mathematics, physics, engineering, astronomy and even land surveying. The properties of numbers fascinated him, and in his work he anticipated major mathematical development of later times. Along with Wilhelm Welser, he built the first electric telegraph in 1833. His mathematical theory of curved surface laid the foundation for the later work of Riemann.

The situation is radically different for magnetic fields which are continuous and form closed loops. Examine the Gaussian surfaces represented by i or ii in Fig 8.3(a) or Fig. 8.3(b). Both cases visually demonstrate that the number of magnetic field lines leaving the surface is balanced by the number of lines entering it. The *net magnetic flux is zero for both the surfaces*. This is true for any closed surface.

Consider a small vector area element $\Delta\mathbf{S}$ of a closed surface S as in Fig. 8.6. The magnetic flux through **ÄS** is defined as $\Delta\phi_B = \mathbf{B}.\Delta\mathbf{S}$, where **B** is the field at $\Delta\mathbf{S}$. We divide **S**into many small area elements and calculate the individual flux through each. Then, the net flux $\phi_B$is,

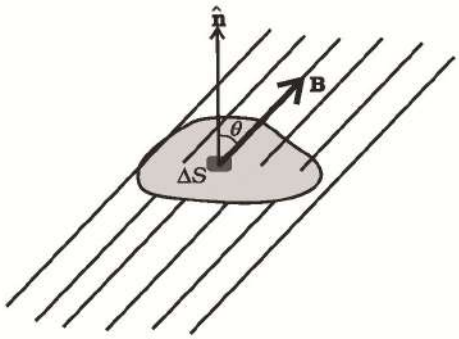


Fig. 8.6

$$\phi_B = \sum_{'all'} \Delta\varphi_B = \sum_{'all'} B.\Delta S = 0 \qquad (8.9)$$

where 'all' stands for 'all area elements $\Delta\mathbf{S}$'. Compare thiswith the Gauss's law of electrostatics. The flux through a closed surface in that case is given by

$$\sum \mathbf{E}.\Delta\mathbf{S} = \frac{q}{\varepsilon_0}$$

where q is the electric charge enclosed by the surface.

The difference between the Gauss's law of magnetism and that for electrostatics is a reflection of the fact that isolated magnetic poles (also called monopoles) are not known to exist. There are no sources or sinks of **B**; the simplest magnetic element is a dipole or a current loop. All magnetic phenomena can be explained in terms of an arrangement of dipoles and/or current loops.

Thus, Gauss's law for magnetism is:

The net magnetic flux through any closed surface is zero.

8.4 THE EARTH'S MAGNETISM

Earlier we have referred to the magnetic field of the earth. The strength of the earth's magnetic field varies from place to place on the earth's surface; its value being of the order of 10^{-5} T.

What causes the earth to have a magnetic field is not clear. Originally the magnetic field was thought of as arising from a giant bar magnet placed approximately along the axis of rotation of the earth and deep in the interior. However, this simplistic picture is certainly not correct. The magnetic field is now thought to arise due to electrical currents produced by convective motion

of metallic fluids (consisting mostly of molten iron and nickel) in the outer core of the earth. This is known as the *dynamo effect*.

The magnetic field lines of the earth resemble that of a (hypothetical) magnetic dipole located at the centre of the earth. The axis of the dipole does not coincide with the axis of rotation of the earth but is presently titled by approximately 11.3° with respect to the later. In this way of looking at it, the magnetic poles are located where the magnetic field lines due to the dipole enter or leave the earth. The location of the north magnetic pole is at a latitude of 79.74° N and a longitude of 71.8° W, a place somewhere in north Canada. The magnetic south pole is at 79.74° S, 108.22° E in the Antarctica.

The pole near the geographic north pole of the earth is called the *north magnetic pole*. Likewise, the pole near the geographic south pole is called the *south magnetic pole*. There is some confusion in the nomenclature of the poles. If one looks at the magnetic field lines of the earth (Fig. 8.8), one sees that unlike in the case of a bar magnet, the field lines go into the earth at the north magnetic pole (*Nm*) and come out from the south magnetic pole (*Sm*). The convention arose because the magnetic north was the direction to which the north pole of a magnetic needle pointed; the north pole of a magnet was so named as it was the *north seeking pole*. Thus, in reality, the north magnetic pole behaves like the south pole of a bar magnet inside the earth and vice versa.

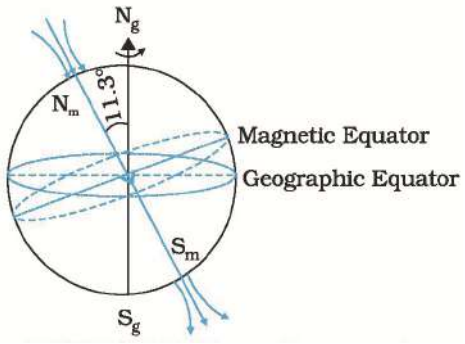


Fig. 8.8 The earth as a giant magnetic dipole.

8.4.1 Magnetic declination and dip

Consider a point on the earth's surface. At such a point, the direction of the longitude circle determines the geographic north-south direction, the line of longitude towards the north pole being the direction of true north. The vertical plane containing the longitude circle and the axis of rotation of the earth is called the *geographic meridian*. In a similar way, one can define *magnetic meridian* of a place as the vertical plane which passes through the imaginary line joining the magnetic north and the south poles. This plane would intersect the surface of the earth in a longitude like circle. A magnetic needle, which is free to swing horizontally, would then lie in the magnetic meridian and the north pole of the needle would point towards the magnetic north pole. Since the line joining the magnetic poles is titled with respect to the geographic axis of the earth, the magnetic meridian at a point makes angle with the geographic meridian. This, then, is the angle between the true geographic north and the north shown by a compass needle. This angle is called the *magnetic declination* or simply *declination* (Fig. 8.9).

There is one more quantity of interest. If a magnetic needle is perfectly balanced about a horizontal axis so that it can swing in a plane of the magnetic meridian, the needle would make an angle with the horizontal (Fig. 8.10). This is known as the *angle of dip* (also known as *inclination*). Thus, dip is the angle that the total magnetic field $\mathbf{B}_E$ of the earth makes with the surface of the earth. Figure 8.11 shows the magnetic meridian plane at a point P on the surface of the earth. The plane is a section through the earth. The total magnetic field at P can be resolved into a horizontal component $\mathbf{H}_E$ and a vertical component $\mathbf{Z}_E$. The angle that $\mathbf{B}_E$ makes with $\mathbf{H}_E$ is the angle of dip,*I*.

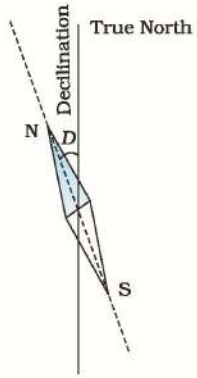


FIGURE 8.9 A magnetic needle free to move in horizontal plane, points toward the magnetic north-south direction.

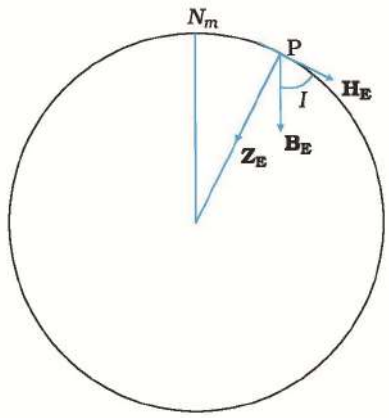


FIGURE 8.10 The circle is a section through the earth containing the magnetic meridian. The angle between $\mathbf{B}_E$ and the horizontal component $\mathbf{H}_E$ is the angle of dip.

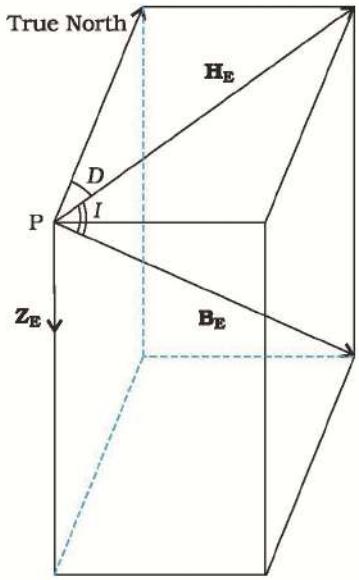


FIGURE 8.11 The earth's magnetic field, $\mathbf{B}_E$, its horizontal and vertical components, $\mathbf{H}_E$ and $\mathbf{Z}_E$. Also shown are the declination, *D* and the inclination or angle of dip, *I*.

In most of the northern hemisphere, the north pole of the dip needle tilts downwards. Likewise in most of the southern hemisphere, the south pole of the dip needle tilts downwards.

To describe the magnetic field of the earth at a point on its surface, we need to specify three quantities, viz., the declination *D*, the angle of dip or the inclination *I* and the horizontal component of the earth's field *H*E. These are known as the *element of the earth's magnetic field*.

Representing the verticle component by Z_E, we have

$$Z_E = B_E \sin I \qquad [8.10(a)]$$

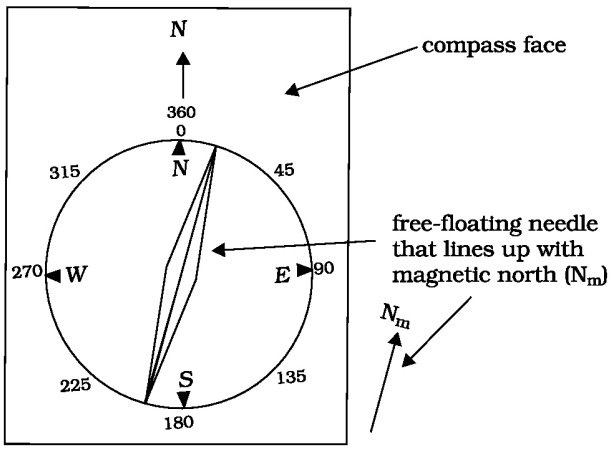
$$H_E = B_E \cos I \qquad [8.10(b)]$$

which gives,

$$\tan I = \frac{Z_E}{H_E} \qquad [8.10(c)]$$

WHAT HAPPENS TO MY COMPASS NEEDLES AT THE POLES?

A compass needle consists of a magnetic needle which floats on a pivotal point. When the compass is held level, it points along the direction of the horizontal component of the earth's magnetic field at the location. Thus, the compass needle would stay along the magnetic meridian of the place. In some places on the earth there are deposits of magnetic minerals which cause the compass needle to deviate from the magnetic meridian. Knowing the magnetic declination at a place allows us to correct the compass to determine the direction of true north.



Diagram

So what happens if we take our compass to the magnetic pole? At the poles, the magnetic field lines are converging or diverging vertically so that the horizontal component is negligible. If the needle is only capable of moving in a horizontal plane, it can point along any direction, rendering it useless as a direction finder. What one needs in such a case is a *dip needle* which is a compass pivoted to move in a vertical plane containing the magnetic field of the earth. The needle of the compass then shows the angle which the magnetic field makes with the vertical. At the magnetic poles such a needle will point straight down.

EARTH'S MAGNETIC FIELD

It must not be assumed that there is a giant bar magnet deep inside the earth which is causing the earth's magnetic field. Although there are large deposits of iron inside the earth, it is highly unlikely that a large solid block of iron stretches from the magnetic north pole to the magnetic south pole. The earth's core is very hot and molten, and the ions of iron and nickel are responsible for earth's magnetism. This hypothesis seems very probable. Moon, which has no molten core, has no magnetic field, Venus has a slower rate of rotation, and a weaker magnetic field, while Jupiter, which has the fastest rotation rate among planets, has a fairly strong magnetic field. However, the precise mode of these circulating currents and the energy needed to sustain them are not very well understood. These are several open questions which form an important area of continuing research.

The variation of the earth's magnetic field with position is also an interesting area of study. Charged particles emitted by the sun flow towards the earth and beyond, in a stream called the

solar wind. Their motion is affected by the earth's magnetic field, and in turn, they affect the pattern of the earth's magnetic field. The pattern of magnetic field near the poles is quite different from that in other regions of the earth.

The variation of earth's magnetic field with time is no less fascinating. There are short term variations taking place over centuries and long term variations taking place over a period of a million years. In a span of 240 years from 1580 to 1820 AD, over which records are available, the magnetic declination at London has been found to change by 3.5°, suggesting that the magnetic poles inside the earth change position with time. On the scale of a million years, the earth's magnetic fields has been found to reverse its direction. Basalt contains iron, and basalt is emitted during volcanic activity. The little iron magnets inside it align themselves parallel to the magnetic field at that place as the basalt cools and solidifies. Geological studies of basalt containing such pieces of magnetised region have provided evidence for the change of direction of earth's magnetic field, several times in the past.

8.5 MAGNETISATION AND MAGNETIC INTENSITY

The earth abounds with a bewildering variety of elements and compounds. In addition, we have been synthesising new alloys, compounds and even elements. One would like to classify the magnetic properties of these substances. In the present section, we define and explain certain terms which will help us to carry out this exercise.

We have seen that a circulating electron in an atom has a magnetic moment. In a bulk material, these moments add up vectorially and they can give a net magnetic moment which is non-zero. We define *magnetisation* **M** of a sample to be equal to its net magnetic moment per unit volume:

$$\mathbf{M} = \frac{m_{net}}{V} \tag{8.11}$$

M is a vector with dimensions L^{-1} A and is measured in a units of A m^{-1}.

Consider a long solenoid of *n* turns per unit length and carrying a current *I*. The magnetic field in the interior of the solenoid was shown to be given by

$$\mathbf{B}_0 = \mu_0 nI \tag{8.12}$$

If the interior of the solenoid is filled with a material with non-zero magnetisation, the field inside the solenoid will be greater than $\mathbf{B}_0$. The net **B** field in the interior of the solenoid may be expressed as

$$\mathbf{B} = \mathbf{B}_0 + \mathbf{B}_m \tag{8.13}$$

where $\mathbf{B}_m$ is the field contributed by the material core. It turns out that this additional field $\mathbf{B}_m$ is proportional to the magnetisation **M** of the material and is expressed as

$$\mathbf{B}_m = \mu_0 \mathbf{M} \tag{8.14}$$

where μ_0 is the same constant (permittivity of vacuum) that appears in Biot-Savart's law.

It is convenient to introduce another vector field **H**, called the *magnetic intensity*, which is defined by

$$\mathbf{H} = \frac{B}{\mu_0} - \mathbf{M} \tag{8.15}$$

where **H** has the same dimensions as **M** and is measured in units of A m^{-1}. Thus, the total magnetic field **B** is written as

$$\mathbf{B} = \mu_0 (\mathbf{H} + \mathbf{M}) \quad (8.16)$$

We repeat our defining procedure. We have partitioned the contribution to the total magnetic field inside the sample into two parts: *one*, due to external factors such as the current in the solenoid. This is represented by **H**. The *other* is due to the specific nature of the magnetic material, namely **M**. The latter quantity can be influenced by external factors. This influence is mathematically expressed as

$$\mathbf{M} = \chi \mathbf{H} \quad (8.17)$$

where χ , a dimensionless quantity, is appropriately called the *magnetic susceptibility*. It is a measure of how a magnetic material responds to an external field. Table 8.2 lists χ for some elements. It is small and positive for materials, which are called *paramagnetic*. It is small and negative for materials, which are termed *diamagnetic*. In the latter case **M** and **H** are opposite in direction. From Eqs. (8.16) and (8.17) we obtain,

$$\mathbf{B} = \mu_0 (1 + \chi)\mathbf{H} \quad (8.18)$$

$$= \mu_0 \mu_r \mathbf{H}$$

$$= \mu \mathbf{H} \quad (8.19)$$

where $\mu = 1 + \chi$, is a dimensionless quantity called the *relative magnetic permeability* of the substance. It is the analog of the dielectric constant in electrostatics. The *magnetic permeability* of the substance is $\mu$and it has the same dimensions and units as μ_0;

$$\mu = \mu_0 \mu_r = \mu_0 (1+\chi).$$

The three quantities χ, μ and μ are interrelated and only one of them is independent. Given one, the other two may be easily determined.

TABLE 8.2 MAGNETIC SUSCEPTIBILITY OF SOME ELEMENTS AT 300 K

Diamagnetic substance	χ	**Paramagnetic substance**	χ
Bismuth	-1.66×10^{-5}	Aluminium	2.3×10^{-5}
Copper	-9.8×10^{-6}	Calcium	1.9×10^{-5}
Diamond	-2.2×10^{-5}	Chromium	2.7×10^{-4}
Gold	-3.6×10^{-5}	Lithium	2.1×10^{-5}
Lead	-1.7×10^{-5}	Magnesium	1.2×10^{-5}
Mercury	-2.9×10^{-5}	Niobium	2.6×10^{-5}
Nitrogen (STP)	-5.0×10^{-9}	Oxygen (STP)	2.1×10^{-6}
Silver	-2.6×10^{-5}	Platinum	2.9×10^{-4}
Silicon	-4.2×10^{-6}	Tungsten	6.8×10^{-5}

8.6 MAGNETIC PROPERTIES OF MATERIALS

The discussion in the previous section helps us to classify materials as diamagnetic, paramagnetic or ferromagnetic. In terms of the susceptibility χ , a material is diamagnetic if χ is negative, para- if χ is positive and small, and ferro- if χ is large and positive.

A glance at Table 8.3 gives one a better feeling for these materials. Here $\varepsilon$is a small positive number introduced to quantify paramagnetic materials. Next, we describe these materials in some detail.

TABLE 8.3

Diamagnetic	Paramagnetic	Ferromagnetic
$-1 \leq \chi < 0$	$0 < \chi < \varepsilon$	$\chi >> 1$
$0 \leq \mu_r < 1$	$1 < \mu_r < 1+\varepsilon$	$\mu_r >> 1$
$\mu < \mu_0$	$\mu > \mu_0$	$\mu >> \mu_0$

8.6.1 Diamagnetism

Diamagnetic substances are those which have tendency to move from stronger to the weaker part of the external magnetic field. In other words, unlike the way a magnet attracts metals like iron, it would repel a diamagnetic substance.

Figure 8.12(a) shows a bar of diamagnetic material placed in an external magnetic field. The field lines are repelled or expelled and the field inside the material is reduced. In most cases, as is evident from Table 8.2, this reduction is slight, being one part in 10^5. When placed in a non-uniform magnetic field, the bar will tend to move from high to low field.

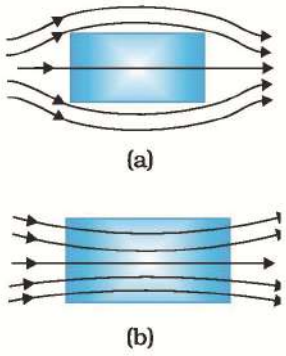


Fig. 8.12 Behaviour of magnetic field lines near a (a)diamagnetic, (b) paramagnetic substance.

The simplest explanation for diamagnetism is as follows. Electrons in an atom orbiting around nucleus possess orbital angular momentum. These orbiting electrons are equivalent to current-carrying loop and thus possess orbital magnetic moment. Diamagnetic substances are the ones in which resultant magnetic moment in an atom is zero. When magnetic field is applied, those electrons having orbital magnetic moment in the same direction slow down and those in the opposite direction speed up. This happens due to induced current in accordance with Lenz's law which you will study in Chapter 9. Thus, the substance develops a net magnetic moment in direction opposite to that of the applied field and hence repulsion.

Some diamagnetic materials are bismuth, copper, lead, silicon, nitrogen (at STP), water and sodium chloride. Diamagnetism is present in all the substances. However, the effect is so weak in most cases that it gets shifted by other effects like paramagnetism, ferromagnetism, etc.

The most exotic diamagnetic materials are *superconductors*. These are metals, cooled to very low temperatures which exhibit both *perfect conductivity* and *perfect diamagnetism*. Here the field lines are completely expelled! $\chi = -1$ and $\mu_r = 0$. A superconductor repels a magnet and (by Newton's third law) is repelled by the magnet. The phenomenon of perfect diamagnetism in superconductors is called the *Meissner effect*, after the name of its discoverer. Superconducting magnets can be gainfully exploited in variety of situations, for example, for running magnetically levitated superfast trains.

8.6.2 Paramagnetism

Paramagnetic substances are those which get weakly magnetised when placed in an external magnetic field. They have tendency to move from a region of weak magnetic field to strong magnetic field, i.e., they get weakly attracted to a magnet.

The individual atoms (or ions or molecules) of a paramagnetic material possess a permanent magnetic dipole moment of their own. On account of the ceaseless random thermal motion of the atoms, no net magnetisation is seen. In the presence of an external field $\mathbf{B}_0$, which

is strong enough, and at low temperatures, the individual atomic dipole moment can be made to align and point in the same direction as $\mathbf{B}_0$. Figure 8.12(b) shows a bar of paramagnetic material placed in an external field. The field lines gets concentrated inside the material, and the field inside is enhanced. In most cases, as is evident from Table 8.2, this enhancement is slight, being one part in 10^5. When placed in a non-uniform magnetic field, the bar will tend to move from weak field to strong.

Some paramagnetic materials are aluminium, sodium, calcium, oxygen (at STP) and copper chloride. Experimentally, one finds that the magnetisation of a paramagnetic material is inversely proportional to the absolute temperature T,

$$M = C\frac{B_0}{T} \qquad [8.20(a)]$$

or equivalently, using Eqs. (8.12) and (8.17)

$$\chi = C\frac{\mu_0}{T} \qquad [8.20(b)]$$

This is known as *Curie's law*, after its discoverer Pieree Curie (1859- 1906). The constant C is called *Curie's constant*. Thus, for a paramagnetic material both χ and μr depend not only on the material, but also (in a simple fashion) on the sample temperature. As the field is increased or the temperature is lowered, the magnetisation increases until it reaches the saturation value *Ms*, at which point all the dipoles are perfectly aligned with the field. Beyond this, Curie's law [Eq. (8.20)] is no longer valid.

8.6.3 Ferromagnetism

Ferromagnetic substances are those which gets strongly magnetised when placed in an external magnetic field. They have strong tendency to move from a region of weak magnetic field to strong magnetic field, i.e., they get strongly attracted to a magnet.

The individual atoms (or ions or molecules) in a ferromagnetic material possess a dipole moment as in a paramagnetic material. However, they interact with one another in such a way that they spontaneously align themselves in a common direction over a macroscopic volume called *domain*. The explanation of this cooperative effect requires quantum mechanics and is beyond the scope of this textbook. Each domain has a net magnetisation. Typical domain size is 1mm and the domain contains about 10^{11} atoms. In the first instant, the magnetisation varies randomly from domain to domain and there is no bulk magnetisation. This is shown in Fig. 8.13(a). When we apply an external magnetic field $\mathbf{B}_0$, the domains orient themselves in the direction of $\mathbf{B}_0$ and simultaneously the domain oriented in the direction of $\mathbf{B}_0$ grow in size. This existence of domains and their motion in $\mathbf{B}_0$ are not speculations. One may observe this under a microscope after sprinkling a liquid suspension of powdered ferromagnetic substance of samples. This motion of suspension can be observed. Figure 8.12(b) shows the situation when the domains have aligned and amalgamated to form a single 'giant' domain.

Thus, in a ferromagnetic material the field lines are highly concentrated. In non-uniform magnetic field, the sample tends to move towards the region of high field. We may wonder as to what happens when the external field is removed. In some ferromagnetic materials the magnetisation persists. Such materials are called *hard* magnetic materials or *hard ferromagnets*. Alnico, an alloy of iron, aluminium, nickel, cobalt and copper, is one such material. The naturally occurring lodestone is another. Such materials form permanent magnets to be used among other things as a compass needle. On the other hand, there is a class of ferromagnetic materials in which the magnetisation disappears on removal of the external field. Soft iron is one such material. Appropriately enough, such materials are called *soft ferromagnetic materials*. There are a number of elements, which are ferromagnetic: iron, cobalt, nickel, gadolinium, etc. The relative magnetic permeability is >1000!

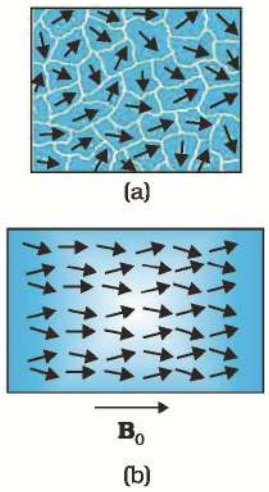


FIGURE 8.13 (a) Randomly oriented domains, (b) Aligned domains.

The ferromagnetic property depends on temperature. At high enough temperature, a ferromagnet becomes a paramagnet. The domain structure disintegrates with temperature. This disappearance of magnetisation with temperature is gradual. It is a phase transition reminding us of the melting of a solid crystal. The temperature of transition from ferromagnetic to paramagnetism is called the *Curie temperature* T_c. Table 8.4 lists the Curie temperature of certain ferromagnets. The susceptibility above the Curie temperature, i.e., in the paramagnetic phase is described by,

$$\chi = \frac{C}{T - T_c} (T > T_c) \qquad (8.21)$$

TABLE 8.4 CURIE TEMPERATURE T_c OF SOME FERROMAGNETIC MATERIALS

Material	*Tc* **(K)**
Cobalt	1394
Iron	1043
Fe_2O_3	893
Nickel	631
Gadolinium	317

SUMMARY

1. The science of magnetism is old. It has been known since ancient times that magnetic materials tend to point in the north-south direction; like magnetic poles repel and unlike ones attract; and cutting a bar magnet in two leads to two smaller magnets. Magnetic poles cannot be isolated.
2. When a bar magnet of dipole moment **m** is placed in a uniform magnetic field **B**,
   (a) the force on it is zero,

(b) the torque on it is **m** × **B**,

(c) its potential energy is **–m.B**, where we choose the zero of energy at the orientation when **m** is perpendicular to **B**.

3. Consider a bar magnet of size *l* and magnetic moment **m**, at a distance *r* from its mid-point, where *r* >>*l*, the magnetic field **B** due to this bar is,

$$\mathbf{B} = \frac{\mu_0 m}{2\pi r^3} \text{ (along axis)}$$

$$= -\frac{\mu_0 m}{4\pi r^3} \qquad \text{(along equator)}$$

4. Gauss's law for magnetism states that the net magnetic flux through any closed surface is zero

$$\phi_B = \sum_{\substack{all\ area \\ elements\ \Delta S}} B.\Delta S = 0$$

5. The earth's magnetic field resembles that of a (hypothetical) magnetic dipole located at the centre of the earth. The pole near the geographic north pole of the earth is called the north magnetic pole. Similarly, the pole near the geographic south pole is called the south magnetic pole. This dipole is aligned making a small angle with the rotation axis of the earth. The magnitude of the field on the earth's surface $\approx 4 \times 10^{-5}$ T.

6. Three quantities are needed to specify the magnetic field of the earth on its surface – the horizontal component, the magnetic declination, and the magnetic dip. These are known as the elements of the earth's magnetic field.

7. Consider a material placed in an external magnetic field $\mathbf{B}_0$. The magnetic intensity is defined as,

$$\mathbf{H} = \frac{B_0}{\mu_0}$$

The magnetisation **M** of the material is its dipole moment per unit volume. The magnetic field **B** in the material is,

$$\mathbf{B} = \mu_0 (\mathbf{H} + \mathbf{M})$$

8. For a linear material $\mathbf{M} = \chi\mathbf{H}$. So that $\mathbf{B} = \mu\mathbf{H}$ and χ is called the magnetic susceptibility of the material. The three quantities, χ, the relative magnetic permeability μ , and the magnetic permeability μ are related as follows:

$$\mu = \mu_0\mu$$

$$\mu_r = 1 + \chi$$

9. Magnetic materials are broadly classified as: diamagnetic, paramagnetic, and ferromagnetic. For diamagnetic materials $\chi$is negative and small and for paramagnetic materials it is positive and small. Ferromagnetic materials have large $\chi$and are

characterised by non-linear relation between **B** and **H**. They show the property of hysteresis.

10. Substances, which at room temperature, retain their ferromagnetic property for a long period of time are called permanent magnets.

Physical quantity	**Symbol**	**Nature**	**Dimensions**	**Units**	**Remarks**
Permeability of free space	μ_0	Scalar	$[MLT^{-2}A^{-2}]$	T m A^{-1}	$\mu_0/4\pi = 10^{-7}$
Magnetic field, Magnetic induction, Magnetic flux density	**B**	Vector	$[MT^{-2}A^{-1}]$	T (tesla)	10^4 G (gauss)= 1 T
Magnetic moment	**m**	Vector	$[L^{-2}A]$	A m^2	
Magnetic flux	ϕ_B	Scalar	$[ML^2T^{-2}A^{-1}]$	W(weber)	W = T m^2
Magnetisation	**M**	Vector	$[L^{-1}A]$	A m^{-1}	$\frac{\text{Magnetic moment}}{\text{Volume}}$
Magnetic intensity Magnetic field strength	**H**	Vector	$[L^{-1}A]$	A m^{-1}	$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M})$
Magnetic susceptibility	χ	Scalar	-	-	$\mathbf{M} = \chi\mathbf{H}$
Relative magnetic permeability	μ_r	Scalar	-	-	$\mathbf{B} = \mu_0\mu_r\,\mathbf{H}$
Magnetic permeability	μ	Scalar	$[MLT^{-2}A^{-2}]$	T m A^{-1} N A^{-2}	$\mu = \mu_0\mu_r$ $\mathbf{B} = \mu\mathbf{H}$

VERY SHORT ANSWER QUESTIONS (2 MARKS)

1. What happens to compass needles at the Earth's poles?
2. What is the magnetic moment associated with a solenoid?
3. What are the units of magnetic moment, magnetic induction and magnetic field?
4. Magnetic lines form continuous closed loops. Why?
5. Define magnetic declination.
6. Define magnetic inclination or angle of dip.
7. Classify the following materials with regard to magnetism: Manganese, Cobalt, Nickel, Bismuth, Oxygen, and Copper.

SHORT ANSWER QUESTIONS (4 MARKS)

1. Compare the properties of para, dia and ferromagnetic substances.
2. Define retentivity and coercivity. Draw the hysteresis curve for soft iron and steel. What do you infer from these curves?

LONG ANSWER QUESTIONS (8 MARKS)

1. Derive an expression for the magnetic field at a point on the axis of a current carrying circular loop.
2. Define magnetic susceptibility of a material. Name two elements one having positive susceptibility and other having negative susceptibility.
3. What do you understand by "hysteresis"? How does this property influence the choice of materials used in different appliances where electromagnets are used?

CHAPTER 9

ELECTRO MAGNETIC INDUCTION

9.1 INTRODUCTION

Electricity and magnetism were considered separate and unrelated phenomena for a long time. In the early decades of the nineteenth century, experiments on electric current by Oersted, Ampere and a few others established the fact that electricity and magnetism are inter-related. They found that moving electric charges produce magnetic fields. For example, an electric current deflects a magnetic compass needle placed in its vicinity. This naturally raises the questions like: Is the converse effect possible? Can moving magnets produce electric currents? Does the nature permit such a relation between electricity and magnetism? The answer is resounding yes! The experiments of Michael Faraday in England and Joseph Henry in USA, conducted around 1830, demonstrated conclusively that electric currents were induced in closed coils when subjected to changing magnetic fields. In this chapter, we will study the phenomena associated with changing magnetic fields and understand the underlying principles. The phenomenon in which electric current is generated by varying magnetic fields is appropriately called *electromagnetic induction*.

When Faraday first made public his discovery that relative motion between a bar magnet and a wire loop produced a small current in the latter, he was asked, "What is the use of it?" His reply was: "What is the use of a new born baby?" The phenomenon of electromagnetic induction is not merely of theoretical or academic interest but also of practical utility. Imagine a world where there is no electricity – no electric lights, no trains, no telephones and no personal computers. The pioneering experiments of Faraday and Henry have led directly to the development of modern day generators and transformers. Today's civilisation owes its progress to a great extent to the discovery of electromagnetic induction.

9.2 THE EXPERIMENTS OF FARADAY AND HENRY

The discovery and understanding of electromagnetic induction are based on a long series of experiments carried out by Faraday and Henry. We shall now describe some of these experiments.

Experiment 9.1

Figure 9.1 shows a coil C_1***** connected to a galvanometer G. When the North-pole of a bar magnet is pushed towards the coil, the pointer in the galvanometer deflects, indicating the presence of electric current in the coil. The deflection lasts as long as the bar magnet is in motion. The galvanometer does not show any deflection when the magnet is held stationary. When the magnet is pulled away from the coil, the galvanometer shows deflection in the opposite direction, which indicates reversal of the current's direction. Moreover, when the South-pole of the bar magnet is moved towards or away from the coil, the deflections in the galvanometer are opposite to that observed with the North-pole for similar movements. Further, the deflection (and hence current) is found to be larger when the magnet is pushed towards or pulled away from the coil faster. Instead, when the bar magnet is held fixed and the coil C_1 is moved towards or away from the magnet, the same effects are observed. It shows that *it is the relative motion between the magnet and the coil that is responsible for generation (induction) of electric current in the coil*.

JOSEPH HENRY (1797 – 1878)

Josheph Henry [1797 – 1878] American experimental physicist, professor at Princeton University and first director of the Smithsonian Institution. He made important improvements in electro- magnets by winding coils of insulated wire around iron pole pieces and invented an electromagnetic motor and a new, efficient telegraph. He discovered self-induction and investigated how currents in one circuit induce currents in another.

***** Wherever the term 'coil' or 'loop' is used, it is assumed that they are made up of conducting material and are prepared using wires which are coated with insulating material.

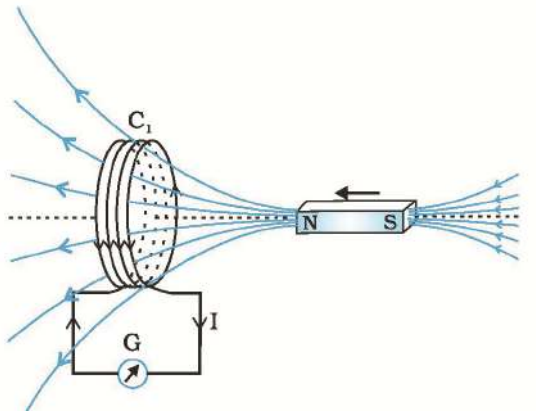


Fig. 9.1 When the bar magnet is pushed towards the coil, the pointer in the galvanometer G deflects.

Experiment 9.2

In Fig. 9.2 the bar magnet is replaced by a second coil C_2 connected to a battery. The steady current in the coil C_2 produces a steady magnetic field. As coil C_2 is moved towards the coil C_1, the galvanometer shows a deflection. This indicates that electric current is induced in coil C_1. When C_2 is moved away, the galvanometer shows a deflection again, but this time in the opposite direction. The deflection lasts as long as coil C_2 is in motion. When the coil C_2 is held fixed and C_1 is moved, the same effects are observed. Again, *it is the relative motion between the coils that induces the electric current*.

Experiment 9.3

The above two experiments involved relative motion between a magnet and a coil and between two coils, respectively. Through another experiment, Faraday showed that this relative motion is not an absolute requirement. Figure 9.3 shows two coils C_1 and C_2 held stationary. Coil C_1 is connected to galvanometer G while the second coil C_2 is connected to a battery through a tapping key K.

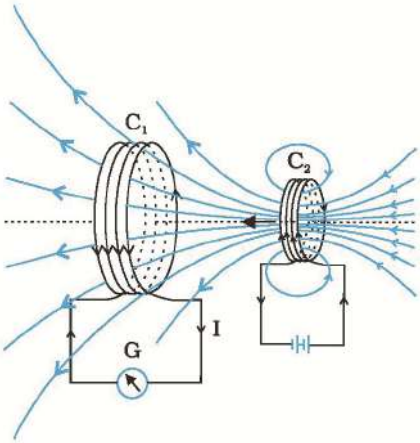


Fig. 9.2 Current is induced in coil C_1 due to motion of the current carrying coil C_2.

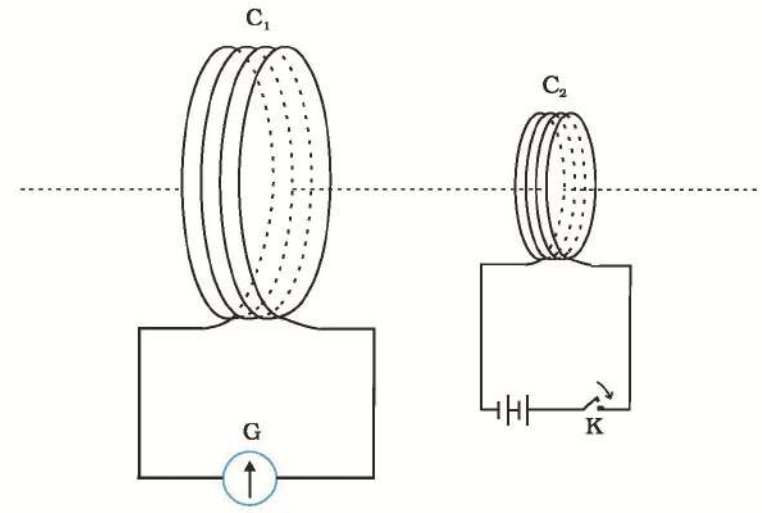


Fig. 9.3 Experimental set-up for Experiment 9.3.

It is observed that the galvanometer shows a momentary deflection when the tapping key K is pressed. The pointer in the galvanometer returns to zero immediately. If the key is held pressed continuously, there is no deflection in the galvanometer. When the key is released, a momentory deflection is observed again, but in the opposite direction. It is also observed that the deflection increases dramatically when an iron rod is inserted into the coils along their axis.

9.3 MAGNETIC FLUX

Faraday's great insight lay in discovering a simple mathematical relation to explain the series of experiments he carried out on electromagnetic induction. However, before we state and appreciate his laws, we must get familiar with the notion of magnetic flux, Φ_B. Magnetic flux is defined in the same way as electric flux is defined in Chapter 1. Magnetic flux through a plane of area A placed in a uniform magnetic field **B** (Fig. 9.4) can be written as

$$\Phi_B = \mathbf{B} \cdot \mathbf{A} = BA \cos\theta \qquad (9.1)$$

where θ is angle between **B** and **A**. The notion of the area as a vector has been discussed earlier in Chapter 4. Equation (9.1) can be extended to curved surfaces and non-uniform fields.

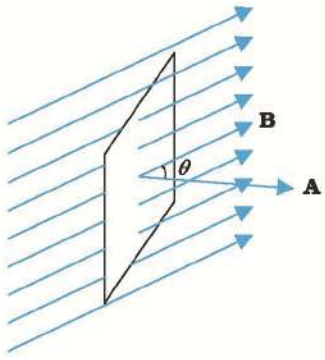


Fig. 9.4 A plane of surface area **A** placed in a uniform magnetic field **B**.

If the magnetic field has different magnitudes and directions at various parts of a surface as shown in Fig. 9.5, then the magnetic flux through the surface is given by

$$\Phi_B = \mathbf{B}_1 . d\mathbf{A}_1 + \mathbf{B}_2 . d\mathbf{A}_2 + \ldots = \sum_{\text{all}} B_i \, . d\mathbf{A}\, i \qquad (9.2)$$

where 'all' stands for summation over all the area elements $d\mathbf{A}_i$ comprising the surface and $\mathbf{B}_i$ is the magnetic field at the area element $d\mathbf{A}_i$. The SI unit of magnetic flux is weber (Wb) or tesla meter squared (T m^2). Magnetic flux is a scalar quantity.

9.4 FARADAY'S LAW OF INDUCTION

From the experimental observations, Faraday arrived at a conclusion that an emf is induced in a coil when magnetic flux through the coil changes with time. Experimental observations discussed in Section 9.2 can be explained using this concept.

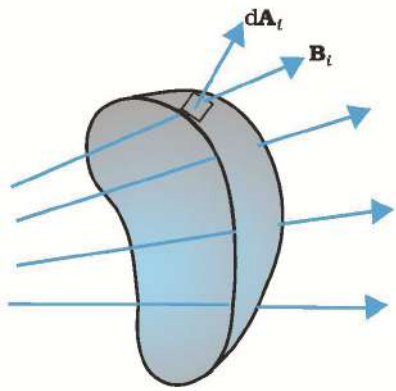


Fig. 9.5 Magnetic field $\mathbf{B}_i$ at the i^{th} area element. d**A** represents area vector of the i^{th} area element.

The motion of a magnet towards or away from coil C_1 in Experiment 9.1 and moving a current-carrying coil C_2 towards or away from coil C_1 in Experiment 9.2, change the magnetic flux associated with coil C_1. The change in magnetic flux induces emf in coil C_1. It was this induced emf which caused electric current to flow in coil C_1 and through the galvanometer. A plausible explanation for the observations of Experiment 9.3 is as follows: When the tapping key K is pressed, the current in coil C_2 (and the resulting magnetic field) rises from zero to a maximum value in a short time. Consequently, the magnetic flux through the neighbouring coil C_1 also increases. It is the change in magnetic flux through coil C_1 that produces an induced emf in coil C_1. When the key is held pressed, current in coil C_2 is constant. Therefore, there is no change in the magnetic flux through coil C_1 and the current in coil C_1 drops to zero. When the key is released, the current in C_2 and the resulting magnetic field decreases from the maximum value to zero in a short time. This results in a decrease in magnetic flux through coil C_1 and hence again induces an electric current in coil C_1*****. The common point in all these observations is that the time rate of change of magnetic flux through a circuit induces emf in it. Faraday stated experimental observations in the form of a law called *Faraday's law of electromagnetic induction*. The law is stated below.

***** Note that sensitive electrical instruments in the vicinity of an electromagnet can be damaged due to the induced emfs (and the resulting currents) when the electromagnet is turned on or off.

The magnitude of the induced emf *in a circuit is equal to the time rate of change of magnetic flux through the circuit.*

Mathematically, the induced emf is given by

$$\varepsilon = -\frac{d\varphi_B}{dt} \tag{9.3}$$

MICHAEL FARADAY (1791–1867)

Michael Faraday [1791–1867] Faraday made numerous contributions to science, viz., the discovery of electromagnetic induction, the laws of electrolysis, benzene, and the fact that the plane of polarisation is rotated in an electric field. He is also credited with the invention of the electric motor, the electric generator and the transformer. He is widely regarded as the greatest experimental scientist of the nineteenth century.

The negative sign indicates the direction of ε and hence the direction of current in a closed loop. This will be discussed in detail in the next section.

In the case of a closely wound coil of *N* turns, change of flux associated with each turn, is the same. Therefore, the expression for the total induced emf is given by

$$\varepsilon = -N\frac{d\varphi_B}{dt} \tag{9.4}$$

The induced emf can be increased by increasing the number of turns *N* of a closed coil.

From Eqs. (9.1) and (9.2), we see that the flux can be varied by changing any one or more of the terms **B**, **A** and θ. In Experiments 9.1 and 9.2 in Section 9.2, the flux is changed by varying **B**. The flux can also be altered by changing the shape of a coil (that is, by shrinking it or stretching it) in a magnetic field, or rotating a coil in a magnetic field such that the angle θ between **B** and **A** changes. In these cases too, an emf is induced in the respective coils.

9.5 LENZ'S LAW AND CONSERVATION OF ENERGY

In 1834, German physicist Heinrich Friedrich Lenz (1804-1865) deduced a rule, known as *Lenz's law* which gives the polarity of the induced emf in a clear and concise fashion. The statement of the law is:

The polarity of induced emf is such that it tends to produce a current which opposes the change in magnetic flux that produced it.

The negative sign shown in Eq. (9.3) represents this effect. We can understand Lenz's law by examining Experiment 9.1 in Section 9.2.1. In Fig. 9.1, we see that the North-pole of a bar magnet is being pushed towards the closed coil. As the North-pole of the bar magnet moves towards the coil, the magnetic flux through the coil increases. Hence current is induced in the coil in such a direction that it opposes the increase in flux. This is possible only if the current in the coil is in a counter-clockwise direction with respect to an observer situated on the side of the magnet. Note that magnetic moment associated with this current has North polarity towards the North-pole of the approaching magnet. Similarly, if the North-pole of the magnet is being withdrawn from the coil, the magnetic flux through the coil will decrease. To counter this decrease in magnetic flux, the induced current in the coil flows in clockwise direction and its South- pole faces the receding North-pole of the bar magnet. This would result in an attractive force which opposes the motion of the magnet and the corresponding decrease in flux.

What will happen if an open circuit is used in place of the closed loop in the above example? In this case too, an emf is induced across the open ends of the circuit. The direction of the induced emf can be found using Lenz's law. Consider Figs. 9.6 (a) and (b). They provide an easier way to understand the direction of induced currents. Note that the direction shown by and indicate the directions of the induced currents.

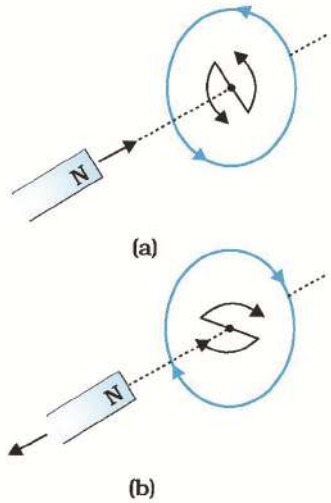


Fig. 9.6 Illustration of Lenz's law.

A little reflection on this matter should convince us on the correctness of Lenz's law. Suppose that the induced current was in the direction opposite to the one depicted in Fig. 6.6(a). In that case, the South-pole due to the induced current will face the approaching North-pole of the magnet. The bar magnet will then be attracted towards the coil at an ever increasing acceleration. A gentle push on the magnet will initiate the process and its velocity and kinetic energy will continuously increase without expending any energy. If this can happen, one could construct a perpetual-motion machine by a suitable arrangement. This violates the law of conservation of energy and hence cannot happen.

Now consider the correct case shown in Fig.9.6(a). In this situation, the bar magnet experiences a repulsive force due to the induced current. Therefore, a person has to do work in moving the magnet.

Where does the energy spent by the person go? This energy is dissipated by Joule heating produced by the induced current.

9.6 MOTIONAL ELECTROMOTIVE FORCE

Let us consider a straight conductor moving in a uniform and time- independent magnetic field. Figure 9.8 shows a rectangular conductor PQRS in which the conductor PQ is free to move. The rod PQ is moved towards the left with a constant velocity **v** as shown in the figure. Assume that there is no loss of energy due to friction. PQRS forms a closed circuit enclosing an area that changes as PQ moves. It is placed in a uniform magnetic field **B** which is perpendicular to the plane of this system. If the length RQ = x and RS = l, the magnetic flux Φ_B enclosed by the loop PQRS will be

$$\Phi_B = Blx$$

Since x is changing with time, the rate of change of flux Φ_B will induce an emf given by:

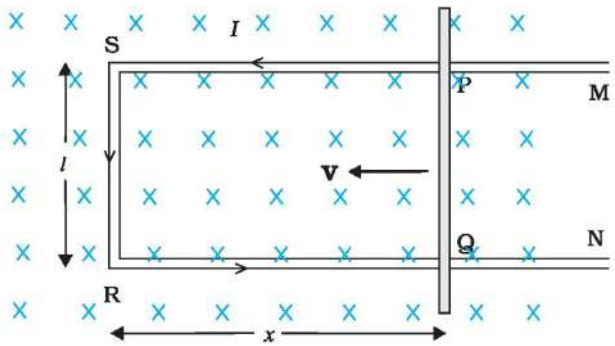


Fig. 9.8 The arm PQ is moved to the left side, thus decreasing the area of the rectangular loop. This movement induces a current I as shown.

$$\varepsilon = \frac{-d\varphi_B}{dt} = -\frac{d}{dt}(Blx)$$

$$= -Bl\frac{dx}{dt} = Blv \qquad (9.5)$$

where we have used $dx/dt = -v$ which is the speed of the conductor PQ. The induced emf Blv is called *motional emf*. Thus, we are able to produce induced emf by moving a conductor instead of varying the magnetic field, that is, by changing the magnetic flux enclosed by the circuit.

It is also possible to explain the motional emf expression in Eq. (9.5) by invoking the Lorentz force acting on the free charge carriers of conductor PQ. Consider any arbitrary charge q in the conductor PQ. When the rod moves with speed v, the charge will also be moving with speed v in the magnetic field **B**. The Lorentz force on this charge is qvB in magnitude, and its direction is towards Q. All charges experience the same force, in magnitude and direction, irrespective of their position in the rod PQ. The work done in moving the charge from P to Q is,

$$W = qvBl$$

Since emf is the work done per unit charge,

$$\varepsilon = \frac{W}{q}$$
$$= Blv$$

This equation gives emf induced across the rod PQ and is identical to Eq. (9.5). We stress that our presentation is not wholly rigorous. But it does help us to understand the basis of Faraday's law when the conductor is moving in a uniform and time-independent magnetic field.

On the other hand, it is not obvious how an emf is induced when a conductor is stationary and the magnetic field is changing – a fact which Faraday verified by numerous experiments. In the case of a stationary conductor, the force on its charges is given by

$$\mathbf{F} = q\,(\mathbf{E} + \mathbf{v} \times \mathbf{B}) = q\mathbf{E} \qquad (9.6)$$

since $\mathbf{v} = 0$. Thus, any force on the charge must arise from the electric field term **E** alone. Therefore, to explain the existence of induced emf or induced current, we must assume that a time-varying magnetic field generates an electric field. However, we hasten to add that electric fields produced by static electric charges have properties different from those produced by time-varying magnetic fields. In Chapter 7, we learnt that charges in motion (current) can exert force/torque on a stationary magnet. Conversely, a bar magnet in motion (or more generally, a changing magnetic field) can exert a force on the stationary charge. This is the fundamental significance of the Faraday's discovery. Electricity and magnetism are related.

9.7 ENERGY CONSIDERATION: A QUANTITATIVE STUDY

In Section 9.5, we discussed qualitatively that Lenz's law is consistent with the law of conservation of energy. Now we shall explore this aspect further with a concrete example.

Let *r* be the resistance of movable arm PQ of the rectangular conductor shown in Fig. 9.10. We assume that the remaining arms QR, RS and SP have negligible resistances compared to *r*. Thus, the overall resistance of the rectangular loop is *r* and this does not change as PQ is moved. The current *I* in the loop is,

$$I = \frac{\varepsilon}{r}$$
$$= \frac{Blv}{r} \qquad (9.7)$$

On account of the presence of the magnetic field, there will be a force on the arm PQ. This force I ($\mathbf{l} \times \mathbf{B}$), is directed outwards in the direction opposite to the velocity of the rod. The magnitude of this force is,

$$F = I\,l\,B = \frac{B^2 l^2 v}{r}$$

where we have used Eq. (9.7). Note that this force arises due to drift velocity of charges (responsible for current) along the rod and the consequent Lorentz force acting on them.

Alternatively, the arm PQ is being pushed with a constant speed *v*, the power required to do this is,

$$P = F\,v$$

$$= \frac{B^2 l^2 v^2}{r} \qquad (9.8)$$

The agent that does this work is mechanical. Where does this mechanical energy go? The answer is: it is dissipated as Joule heat, and is given by

$$P_J = I^2 r = \left(\frac{Blv}{r}\right)^2 r = \frac{B^2 l^2 v^2}{r}$$

which is identical to Eq. (9.8).

Thus, mechanical energy which was needed to move the arm PQ is converted into electrical energy (the induced emf) and then to thermal energy. There is an interesting relationship between the charge flow through the circuit and the change in the magnetic flux. From Faraday's law, we have learnt that the magnitude of the induced emf is,

$$|\varepsilon| = \frac{\Delta\varphi_B}{\Delta t}$$

However,

$$|\varepsilon| = Ir = \frac{\Delta Q}{\Delta t} r$$

Thus,

$$\Delta Q = \frac{\Delta\varphi_B}{r}$$

9.8 EDDY CURRENTS

So far we have studied the electric currents induced in well defined paths in conductors like circular loops. Even when bulk pieces of conductors are subjected to changing magnetic flux, induced currents are produced in them. However, their flow patterns resemble swirling eddies in water. This effect was discovered by physicist Foucault (1819-1868) and these currents are called *eddy currents*.

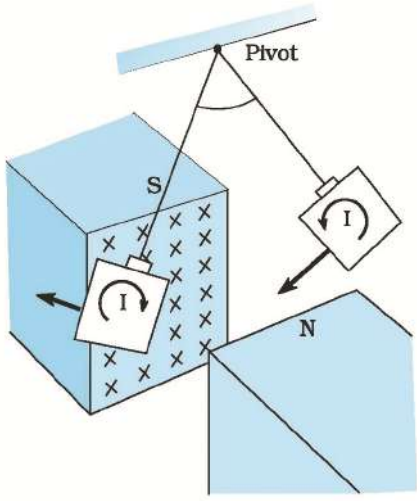


Fig.9.9 Eddy currents are generated in the copper plate, while entering and leaving the region of magnetic field.

Consider the apparatus shown in Fig. 9.9. A copper plate is allowed to swing like a simple pendulum between the pole pieces of a strong magnet. It is found that the motion is damped and in a little while the plate comes to a halt in the magnetic field. We can explain this phenomenon on the basis of electromagnetic induction. Magnetic flux associated with the plate keeps on changing as the plate moves in and out of the region between magnetic poles. The flux change induces eddy currents in the plate. Directions of eddy currents are opposite when the plate swings into the region between the poles and when it swings out of the region.

If rectangular slots are made in the copper plate as shown in Fig. 9.10, area available to the flow of eddy currents is less. Thus, the pendulum plate with holes or slots reduces electromagnetic damping and the plate swings more freely. Note that magnetic moments of the

induced currents (which oppose the motion) depend upon the area enclosed by the currents (recall equation **m** = *I***A** in Chapter 7).

This fact is helpful in reducing eddy currents in the metallic cores of transformers, electric motors and other such devices in which a coil is to be wound over metallic core. Eddy currents are undesirable since they heat up the core and dissipate electrical energy in the form of heat. Eddy currents are minimised by using laminations of metal to make a metal core. The laminations are separated by an insulating material like lacquer. The plane of the laminations must be arranged parallel to the magnetic field, so that they cut across the eddy current paths. This arrangement reduces the strength of the eddy currents. Since the dissipation of electrical energy into heat depends on the square of the strength of electric current, heat loss is substantially reduced.

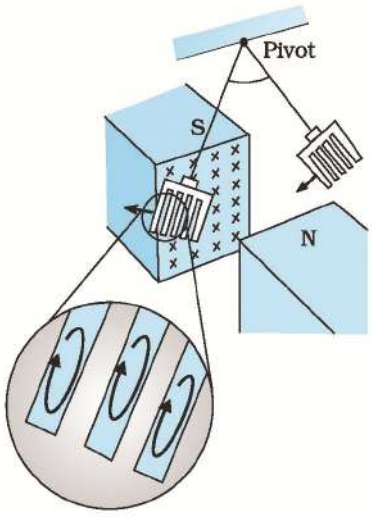


Fig. 9.10 Cutting slots in the copper plate reduces the effect of eddy currents.

Eddy currents are used to advantage in certain applications like:

(i) *Magnetic braking in trains*: Strong electromagnets are situated above the rails in some electrically powered trains. When the electromagnets are activated, the eddy currents induced in the rails oppose the motion of the train. As there are no mechanical linkages, the braking effect is smooth.

(ii) *Electromagnetic damping*: Certain galvanometers have a fixed core made of nonmagnetic metallic material. When the coil oscillates, the eddy currents generated in the core oppose the motion and bring the coil to rest quickly.

(iii) *Induction furnace*: Induction furnace can be used to produce high temperatures and can be utilised to prepare alloys, by melting the constituent metals. A high frequency alternating current is passed through a coil which surrounds the metals to be melted. The eddy currents generated in the metals produce high temperatures sufficient to melt it.

(iv) *Electric power meters*: The shiny metal disc in the electric power meter (analogue type) rotates due to the eddy currents. Electric currents are induced in the disc by magnetic fields produced by sinusoidally varying currents in a coil.

You can observe the rotating shiny disc in the power meter of your house.

ELECTROMAGNETIC DAMPING

Take two hollow thin cylindrical pipes of equal internal diameters made of aluminium and PVC, respectively. Fix them vertically with clamps on retort stands. Take a small cylinderical magnet having diameter slightly smaller than the inner diameter of the pipes and drop it through each pipe in such a way that the magnet does not touch the sides of the pipes during its fall. You will observe that the magnet dropped through the PVC pipe takes the same time to come out of the pipe as it would take when dropped through the same height without the pipe. Note the time it takes to come out of the pipe in each case. You will see that the magnet takes much longer time in the case of aluminium pipe. Why is it so? It is due to the eddy currents that are generated in the aluminium pipe which oppose the change in magnetic flux, i.e., the motion of the magnet. The retarding force due to the eddy currents inhibits the motion of the magnet. Such phenomena

are referred to as *electromagnetic damping*. Note that eddy currents are not generated in PVC pipe as its material is an insulator whereas aluminium is a conductor.

9.9 INDUCTANCE

An electric current can be induced in a coil by flux change produced by another coil in its vicinity or flux change produced by the same coil. These two situations are described separately in the next two sub-sections. However, in both the cases, the flux through a coil is proportional to the current. That is, $\Phi_B \alpha I$.

Further, if the geometry of the coil does not vary with time then,

$$\frac{d\varphi_B}{dt} \propto \frac{dI}{dt}$$

For a closely wound coil of N turns, the same magnetic flux is linked with all the turns. When the flux Φ_B through the coil changes, each turn contributes to the induced emf. Therefore, a term called *flux linkage* is used which is equal to $N\Phi_B$ for a closely wound coil and in such a case

$$N\Phi_B \propto I$$

The constant of proportionality, in this relation, is called *inductance*. We shall see that inductance depends only on the geometry of the coil and intrinsic material properties. This aspect is akin to capacitance which for a parallel plate capacitor depends on the plate area and plate separation (geometry) and the dielectric constant K of the intervening medium (intrinsic material property).

Inductance is a scalar quantity. It has the dimensions of [M L^2 T^{-2} A^{-2}] given by the dimensions of flux divided by the dimensions of current. The SI unit of inductance is *henry* and is denoted by H. It is named in honour of Joseph Henry who discovered electromagnetic induction in USA, independently of Faraday in England.

9.9.1 Mutual inductance

Consider Fig. 9.11 which shows two long co-axial solenoids each of length l. We denote the radius of the inner solenoid S_1 by r_1 and the number of turns per unit length by n_1. The corresponding quantities for the outer solenoid S_2 are r_2 and n_2, respectively. Let N_1 and N_2 be the total number of turns of coils S_1 and S_2, respectively.

When a current I_2 is set up through S_2, it in turn sets up a magnetic flux through S_1. Let us denote it by Φ_1. The corresponding flux linkage with solenoid S_1 is

$$N_1\Phi_1 = M_{12}I_2 \qquad (9.7)$$

M_{12} is called the *mutual inductance* of solenoid S_1 with respect to solenoid S_2. It is also referred to as the *coefficient of mutual induction*.

For these simple co-axial solenoids it is possible to calculate M_{12}. The magnetic field due to the current I_2 in S_2 is $\mu_0 n_2 I_2$. The resulting flux linkage with coil S_1 is,

$$N_1\Phi_1 = (n_1 l)(\pi r_1^2)(\mu_0 n_2 I_2)$$
$$= \mu_0 n_1 n_2 \pi r_1^2 l \qquad (9.8)$$

where $n_1 l$ is the total number of turns in solenoid S_1. Thus, from Eq. (9.7) and Eq. (9.8),

$$M_{12} = \mu_0 n_1 n_2 \pi r_1^2 l \qquad (9.10)$$

Note that we neglected the edge effects and considered the magnetic field $\mu_0 n_2 I_2$ to be uniform throughout the length and width of the solenoid S_2. This is a good approximation keeping in mind that the solenoid is long, implying $l >> r_2$.

We now consider the reverse case. A current I_1 is passed through the solenoid S_1 and the flux linkage with coil S_2 is,

$$N_2\Phi_2 = M_{21} I_1 \qquad (9.10)$$

M_{21} is called the *mutual inductance* of solenoid S_2 with respect to solenoid S_1.

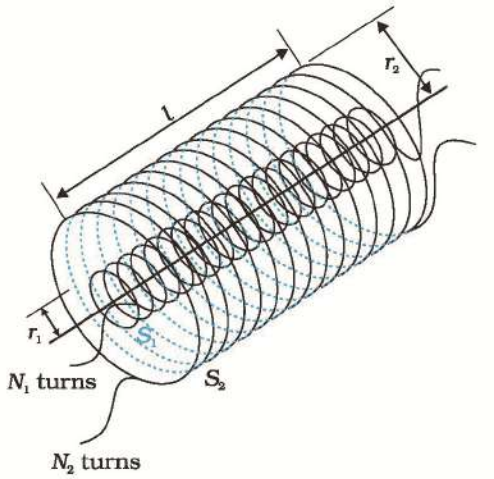


Fig. 9.11 Two long co-axial solenoids of same length *l*.

The flux due to the current I_1 in S_1 can be assumed to be confined solely inside S_1 since the solenoids are very long. Thus, flux linkage with solenoid S_2 is

$$N_2\Phi_2 = (n_2 l)(\pi r_1^2)(\mu_0 n_2 I_2)$$

where $n_2 l$ is the total number of turns of S_2. From Eq. (9.10),

$$M_{21} = \mu_0 n_1 n_2 \pi r_1^2 l \qquad (9.11)$$

Using Eq. (9.9) and Eq. (9.10), we get

$$M_{12} = M_{21} = M \text{ (say)} \qquad (9.12)$$

We have demonstrated this equality for long co-axial solenoids. However, the relation is far more general. Note that if the inner solenoid was much shorter than (and placed well inside) the outer solenoid, then we could still have calculated the flux linkage $N_1\Phi_1$ because the inner solenoid is effectively immersed in a uniform magnetic field due to the outer solenoid. In this case, the calculation of M_{12} would be easy. However, it would be extremely difficult to calculate the flux linkage with the outer solenoid as the magnetic field due to the inner solenoid would vary across the length as well as cross section of the outer solenoid. Therefore, the calculation of M_{21} would also be extremely difficult in this case. The equality $M_{12} = M_{21}$ is very useful in such situations.

We explained the above example with air as the medium within the solenoids. Instead, if a medium of relative permeability μ_r had been present, the mutual inductance would be

$$M = \mu_r \mu_0 n_1 n_2 \pi r_1^2 l$$

It is also important to know that the mutual inductance of a pair of coils, solenoids, etc., depends on their separation as well as their relative orientation.

Now, let us recollect Experiment 9.3 in Section 9.2. In that experiment, emf is induced in coil C_1 wherever there was any change in current through coil C_2. Let Φ_1 be the flux through coil C_1 (say of N_1 turns) when current in coil C_2 is I_2.

Then, from Eq. (9.7), we have

$$N_1\Phi_1 = MI_2$$

For currents varrying with time,

$$\frac{d\left(N_1\varphi_1\right)}{dt} = \frac{d\left(MI_2\right)}{dt}$$

Since induced emf in coil C_1 is given by

$$\varepsilon = -\frac{d\left(N_1\varphi_1\right)}{dt}$$

We get,

$$\varepsilon = -M\frac{dI_2}{dt}$$

It shows that varying current in a coil can induce emf in a neighbouring coil. The magnitude of the induced emf depends upon the rate of change of current and mutual inductance of the two coils.

9.9.2 Self-inductance

In the previous sub-section, we considered the flux in one solenoid due to the current in the other. It is also possible that emf is induced in a single isolated coil due to change of flux through the coil by means of varying the current through the same coil. This phenomenon is called *self-induction*. In this case, flux linkage through a coil of N turns is proportional to the current through the coil and is expressed as

$$N\Phi_B \propto I$$

$$N\Phi_B = L\,I \qquad (9.13)$$

where constant of proportionality L is called *self-inductance* of the coil. It is also called the *coefficient of self-induction* of the coil. When the current is varied, the flux linked with the coil also changes and an emf is induced in the coil. Using Eq. (9.13), the induced emf is given by

$$\varepsilon = -\frac{d\left(N\varphi_B\right)}{dt}$$

$$\varepsilon = -L\frac{dI}{dt} \qquad (9.14)$$

Thus, the self-induced emf always opposes any change (increase or decrease) of current in the coil.

It is possible to calculate the self-inductance for circuits with simple geometries. Let us calculate the self-inductance of a long solenoid of cross- sectional area A and length l, having n turns per unit length. The magnetic field due to a current I flowing in the solenoid is $B = \mu_0 n\,I$ (neglecting edge effects, as before). The total flux linked with the solenoid is

$$N\Phi_B = (nl)(\mu_0 n\,I)\,(A)$$

$$= \mu_0 n^2 Al\,I$$

where nl is the total number of turns. Thus, the self-inductance is,

$$L = \frac{N\varphi_B}{I}$$

$$= \mu_0 n_2 A l \qquad (9.15)$$

If we fill the inside of the solenoid with a material of relative permeability μ_r(for example soft iron, which has a high value of relative permeability), then,

$$L = \mu_r \mu_0 n^2 A l \qquad (9.16)$$

The self-inductance of the coil depends on its geometry and on the permeability of the medium.

The self-induced emf is also called the *back emf* as it opposes any change in the current in a circuit. Physically, the *self-inductance plays the role of inertia*. It is the electromagnetic analogue of mass in mechanics. So, work needs to be done against the back emf (ε) in establishing the current. This work done is stored as magnetic potential energy. For the current I at an instant in a circuit, the rate of work done is

$$\frac{dW}{dt} = |\varepsilon| I$$

If we ignore the resistive losses and consider only inductive effect, then using Eq. (9.14),

$$\frac{dW}{dt} = L I \frac{dI}{dt}$$

Total amount of work done in establishing the current I is

$$W = \int dW = \int_0^1 L I \, dI$$

Thus, the energy required to build up the current I is,

$$W = \frac{1}{2} L I^2 \qquad (9.17)$$

This expression reminds us of $mv^2/2$ for the (mechanical) kinetic energy of a particle of mass m, and shows that L is analogous to m (i.e., L is electrical inertia and opposes growth and decay of current in the circuit).

Consider the general case of currents flowing simultaneously in two nearby coils. The flux linked with one coil will be the sum of two fluxes which exist independently. Equation (9.7) would be modified into

$$N_1 \Phi_1 = M_{11} I_1 + M_{12} I_2$$

where M_{11} represents inductance due to the same coil.

Therefore, using Faraday's law,

$$\varepsilon_1 = -M_{11} \frac{dI_1}{dt} - M_{12} \frac{dI_2}{dt}$$

M_{11} is the *self-inductance* and is written as L_1. Therefore,

$$\varepsilon_1 = -L_1 \frac{dI_1}{dt} - M_{12} \frac{dI_2}{dt}$$

9.10 AC GENERATOR

The phenomenon of electromagnetic induction has been technologically exploited in many ways. An exceptionally important application is the generation of alternating currents (ac). The

modern ac generator with a typical output capacity of 100 MW is a highly evolved machine. In this section, we shall describe the basic principles behind this machine. The Yugoslav inventor Nicola Tesla is credited with the development of the machine. As was pointed out in Section 9.3, one method to induce an emf or current in a loop is through a change in the loop's orientation or a change in its effective area. As the coil rotates in a magnetic field **B**, the effective area of the loop (the face perpendicular to the field) is $A \cos \theta$, where θ is the angle between **A** and **B**. This method of producing a flux change is the principle of operation of a simple ac generator. An ac generator converts mechanical energy into electrical energy.

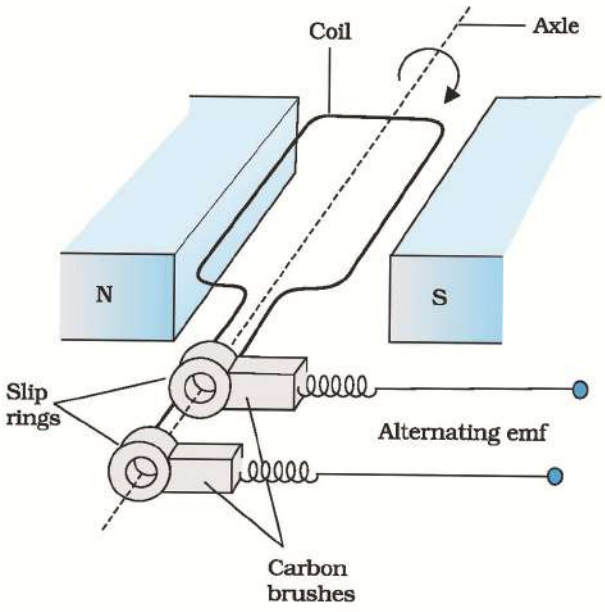


Fig. 9.12 AC Generator

The basic elements of an ac generator are shown in Fig. 9.12. It consists of a coil mounted on a rotor shaft. The axis of rotation of the coil is perpendicular to the direction of the magnetic field. The coil (called armature) is mechanically rotated in the uniform magnetic field by some external means. The rotation of the coil causes the magnetic flux through it to change, so an emf is induced in the coil. The ends of the coil are connected to an external circuit by means of slip rings and brushes.

When the coil is rotated with a constant angular speed ω, the angle θbetween the magnetic field vector **B** and the area vector **A** of the coil at any instant t is $\theta = \omega t$ (assuming $\theta = 0°$ at $t = 0$). As a result, the effective area of the coil exposed to the magnetic field lines changes with time, and from Eq. (9.1), the flux at any time t is

$$\Phi_B = BA \cos \theta = BA \cos \omega t$$

From Faraday's law, the induced emf for the rotating coil of N turns is then,

$$\varepsilon = -N \frac{d\varphi_B}{dt} = -NBA \frac{d}{dt}(\cos \omega t)$$

Thus, the instantaneous value of the emf is

$$\varepsilon = NBA\, \omega \sin \omega t \qquad (9.18)$$

where $NBA\omega$is the maximum value of the emf, which occurs when $\sin \omega t = \pm 1$. If we denote $NBA\omega$as ε_0, then

$$\varepsilon = \varepsilon_0 \sin \omega t \qquad (9.19)$$

Since the value of the sine function varies between +1 and –1, the sign, or polarity of the emf changes with time. Note from Fig. 9.13 that the emf has its extremum value when $\theta = 90°$ or $\theta = 270°$, as the change of flux is greatest at these points.

The direction of the current changes periodically and therefore the current is called *alternating current* (ac). Since $\omega = 2\pi\nu$, Eq (9.19) can be written as

$$\varepsilon = \varepsilon_0 \sin 2\pi\nu t \qquad (9.20)$$

where ν is the frequency of revolution of the generator's coil.

Note that Eq. (9.19) and (9.20) gives the instantaneous value of the emf and ε varies between $+\varepsilon_0$ and $-\varepsilon_0$ periodically. We shall learn how to determine the time-averaged value for the alternating voltage and current in the next chapter.

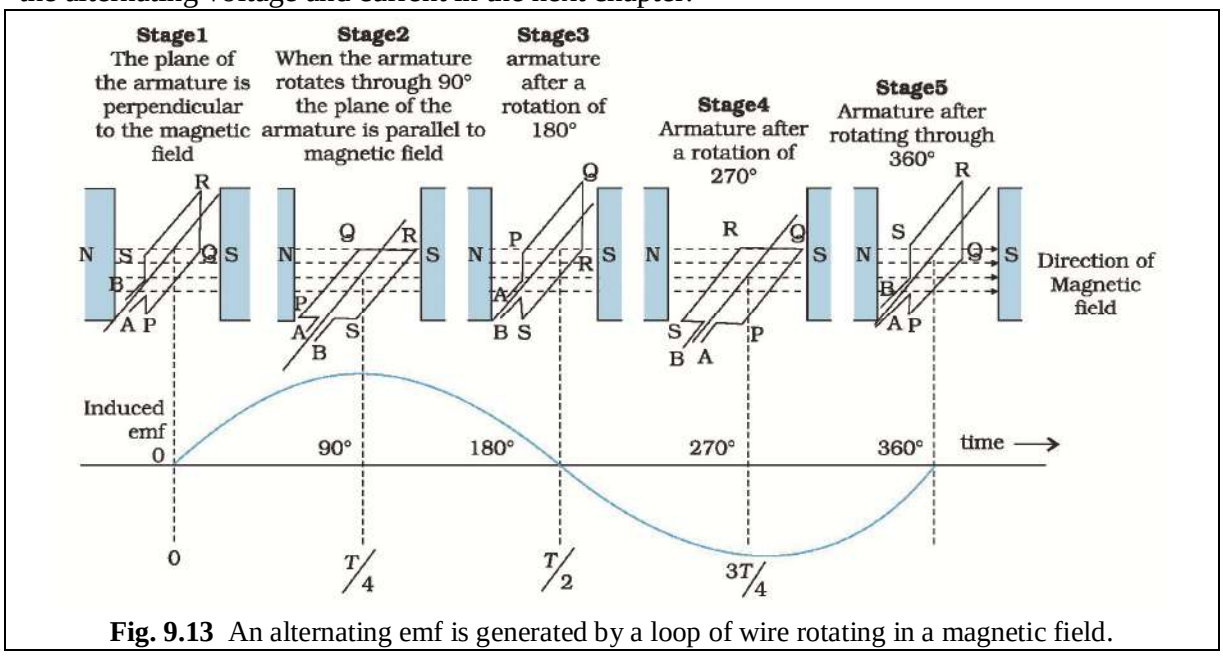


Fig. 9.13 An alternating emf is generated by a loop of wire rotating in a magnetic field.

In commercial generators, the mechanical energy required for rotation of the armature is provided by water falling from a height, for example, from dams. These are called *hydro-electric generators*. Alternatively, water is heated to produce steam using coal or other sources. The steam at high pressure produces the rotation of the armature. These are called *thermal generators*. Instead of coal, if a nuclear fuel is used, we get *nuclear power generators*. Modern day generators produce electric power as high as 500 MW, i.e., one can light up 5 million 100 W bulbs! In most generators, the coils are held stationary and it is the electromagnets which are rotated. The frequency of rotation is 50 Hz in India. In certain countries such as USA, it is 60 Hz.

MIGRATION OF BIRDS

The migratory pattern of birds is one of the mysteries in the field of biology, and indeed all of science. For example, every winter birds from Siberia fly unerringly to water spots in the Indian subcontinent. There has been a suggestion that electromagnetic induction may provide a clue to these migratory patterns. The earth's magnetic field has existed throughout evolutionary history. It would be of great benefit to migratory birds to use this field to determine the direction. As far as we know birds contain no ferromagnetic material so electromagnetic induction seems to be the only reasonable mechanism to determine direction. Consider the optimal case where the magnetic field **B**, the velocity of the bird **v**, and two relevant points of its anatomy separated by a distance *l*, all three are mutually perpendicular. From the formula for motional emf, Eq. (9.5),

$$\varepsilon = Blv$$

Taking $B = 4 \times 10^{-5}$ T, $l = 2$ cm wide, and $v = 10$ m/s, we obtain

$$\varepsilon = 4 \times 10^{-5} \times 2 \times 10^{-2} \times 10 \text{ V} = 8 \times 10^{-6} \text{ V}$$

$$= 8 \ \mu\text{V}$$

This extremely small potential difference suggests that our hypothesis is of doubtful validity. Certain kinds of fish are able to detect small potential differences. However, in these fish, special cells have been identified which detect small voltage differences. In birds no such cells have been identified. Thus, the migration patterns of birds continue to remain a mystery.

SUMMARY

1. The magnetic flux through a surface of area **A** placed in a uniform magnetic field **B** is defined as,

   $\Phi_B = \mathbf{B}.\mathbf{A} = BA \cos\theta$

   where θ is the angle between **B** and **A**.

2. Faraday's laws of induction imply that the emf induced in a coil of N turns is directly related to the rate of change of flux through it,

   $$\varepsilon = -N\frac{d\varphi_B}{dt}$$

   Here Φ_B is the flux linked with one turn of the coil. If the circuit is closed, a current $I = \varepsilon/R$ is set up in it, where R is the resistance of the circuit.

3. Lenz's law states that the polarity of the induced emf is such that it tends to produce a current which opposes the change in magnetic flux that produces it. The negative sign in the expression for Faraday's law indicates this fact.

4. When a metal rod of length l is placed normal to a uniform magnetic field B and moved with a velocity v perpendicular to the field, the induced emf (called motional emf) across its ends is

   $\varepsilon = Bl\,v$

5. Changing magnetic fields can set up current loops in nearby metal (any conductor) bodies. They dissipate electrical energy as heat. Such currents are eddy currents.

6. Inductance is the ratio of the flux-linkage to current. It is equal to $N\Phi/I$.

7. A changing current in a coil (coil 2) can induce an emf in a nearby coil (coil 1). This relation is given by,

   $$\varepsilon_1 = -M_{12}\frac{dI_2}{dt}$$

   The quantity M_{12} is called mutual inductance of coil 1 with respect to coil 2. One can similarly define M_{21}. There exists a general equality,

   $M_{12} = M_{21}$

8. When a current in a coil changes, it induces a back emf in the same coil. The self-induced emf is given by,

   $$\varepsilon = -L\frac{dI}{dt}$$

L is the self-inductance of the coil. It is a measure of the inertia of the coil against the change of current through it.

9. The self-inductance of a long solenoid, the core of which consists of a magnetic material of permeability μ_r, is given by

$L = \mu_r \mu_0 n^2 A l$

where A is the area of cross-section of the solenoid, l its length and n the number of turns per unit length.

10. In an ac generator, mechanical energy is converted to electrical energy by virtue of electromagnetic induction. If coil of N turn and area A is rotated at ν revolutions per second in a uniform magnetic field B, then the motional emf produced is

$\varepsilon = NBA\ (2\pi\nu) \sin(2\pi\nu t)$

where we have assumed that at time t = 0 s, the coil is perpendicular to the field.

Quantity	**Symbol**	**Units**	**Dimensions**	**Equations**
Magnetic Flux	Φ_B	Wb (weber)	$[M L^2 T^{-2} A^{-1}]$	$\Phi_B = \mathbf{B} . \mathbf{A}$
EMF	ε	V (volt)	$[M L^2 T^{-3} A^{-1}]$	$\varepsilon = -\, d(N\Phi_B)/\, dt$
Mutual Inductance	M	H (henry)	$[M L^2 T^{-2} A^{-2}]$	$\varepsilon_1 = -M\, (dI_2/\, dt)$
Self Inductance	L	H (henry)	$[M L^2 T^{-2} A^{-2}]$	$\varepsilon_2 = -L\, (dI / dt)$

VERY SHORT ANSWER QUESTIONS (2 MARKS)

1. What did the experiments of Faraday and Henry show?
2. Define magnetic flux.
3. State Faraday's law of electromagnetic induction.
4. State Lenz's law.
5. What are Eddy currents?
6. Define 'inductance'.
7. What do you understand by 'self inductance'?

SHORT ANSWER QUESTIONS (4 MARKS)

1. Obtain an expression for the emf induced across a conductor which is moved in a uniform magnetic field which is perpendicular to the plane of motion.
2. Describe the ways in which Eddy currents are used to advantage.
3. Obtain an expression for the magnetic energy stored in a soilenoid in terms of the magnetic field, area and length of the solenoid.

LONG ANSWER QUESTIONS (8 MARKS)

1. Outline the path-breaking experiments of Faraday and Henry and highlight the contributions of these experiments to our understanding of electromagnetism.
2. Describe the working of a AC generator with the aid of a simple diagram and necessary expressions.

CHAPTER 10

ALTERNATING CURRENT

10.1 INTRODUCTION

We have so far considered direct current (dc) sources and circuits with dc sources. These currents do not change direction with time. But voltages and currents that vary with time are very common. The electric mains supply in our homes and offices is a voltage that varies like a sine function with time. Such a voltage is called *alternating voltage* (ac voltage) and the current driven by it in a circuit is called the *alternating current* (ac current)*****. Today, most of the electrical devices we use require ac voltage. This is mainly because most of the electrical energy sold by power companies is transmitted and distributed as alternating current. The main reason for preferring use of ac voltage over dc voltage is that ac voltages can be easily and efficiently converted from one voltage to the other by means of transformers. Further, electrical energy can also be transmitted economically over long distances. AC circuits exhibit characteristics which are exploited in many devices of daily use. For example, whenever we tune our radio to a favourite station, we are taking advantage of a special property of ac circuits – one of many that you will study in this chapter.

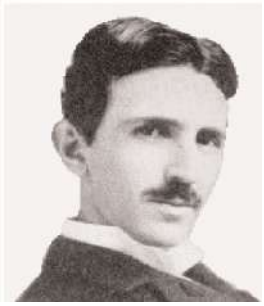
***** The phrases *ac voltage* and *ac current* are contradictory and redundant, respectively, since they mean, literally, *alternating current voltage* and *alternating current current*. Still, the abbreviation *ac* to designate an electrical quantity displaying simple harmonic time dependance has become so universally accepted that we follow others in its use. Further, *voltage* – another phrase commonly used means potential difference between two points.

10.2 AC VOLTAGE APPLIED TO A RESISTOR

Figure 10.1 shows a resistor connected to a source ε of ac voltage. The symbol for an ac source in a circuit diagram is ⏦ . We consider a source which produces sinusoidally varying potential difference across its terminals. Let this potential difference, also called ac voltage, be given by

$$v = v_m \sin \omega t \qquad (10.1)$$

where v_m is the amplitude of the oscillating potential difference and ω is its angular frequency.



NICOLA TESLA (1856 – 1943)

Nicola Tesla (1856–1943) Serbian-American scientist, inventor and genius. He conceived the idea of the rotating magnetic field, which is the basis of practically all alternating current machinery, and which helped usher in the age of electric power. He also invented among other things the induction motor, the polyphase system of ac power, and the high frequency induction coil (the Tesla coil) used in radio and television sets and other electronic equipment. The SI unit of magnetic field is named in his honour.

To find the value of current through the resistor, we apply Kirchhoff's loop rule $\sum\varepsilon(t) = 0$ (refer to Section 3.13), to the circuit shown in Fig. 10.1 to get

$$v_m \sin \omega t = i R$$

$$\text{or} \quad i = \frac{v_m}{R} \sin \omega t$$

Fig. 10.1 AC voltage applied to a resistor.

Since R is a constant, we can write this equation as

$$i = i_m \sin \omega t \qquad (10.2)$$

where the current amplitude i_m is given by

$$i_m = \frac{v_m}{R} \qquad (10.3)$$

Equation (10.3) is Ohm's law, which for resistors, works equally well for both ac and dc voltages. The voltage across a pure resistor and the current through it given by Eqs. (10.1) and (10.2) are plotted as a function of time in Fig. 10.2. Note, in particular that both v and i reach zero, minimum and maximum values at the same time. Clearly, *the voltage* and *current are in phase with each other*.

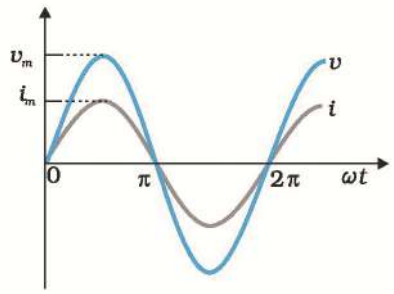


Fig. 10.2 In a pure resistor, the voltage and current are in phase. The minima, zero and maxima occur at the same respective times.

We see that, like the applied voltage, the current varies sinusoidally and has corresponding positive and negative values during each cycle. Thus, the sum of the instantaneous current values over one complete cycle is zero, and the average current is zero. The fact that the average current is zero, however, does not mean that the average power consumed is zero and that there is no dissipation of electrical energy. As you know, Joule heating is given by i^2R and depends on i^2 (which is always positive whether i is positive or negative) and not on i. Thus, there is Joule heating and dissipation of electrical energy when an ac current passes through a resistor.

GEORGE WESTINGHOUSE (1846 – 1914)

George Westinghouse (1846 – 1914) A leading proponent of the use of alternating current over direct current. Thus, he came into conflict with Thomas Alva Edison, an advocate of direct current. Westinghouse was convinced that the technology of alternating current was the key to the electrical future. He founded the famous Company named after him and enlisted the services of Nicola Tesla and other inventors in the development of alternating current motors and apparatus for the transmission of high tension current,

pioneering in large scale lighting.

The instantaneous power dissipated in the resistor is

$$\bar{p} = i^2 R = i_m^2 R \sin^2 \omega t \qquad (10.4)$$

The average value of p over a cycle is*

$$\bar{p} = \langle i^2 R \rangle = \langle i_m^2 R \sin^2 \omega t \rangle \qquad [10.5(a)]$$

where the bar over a letter (here, p) denotes its average value and <.....> denotes taking average of the quantity inside the bracket. Since, i^2 and R are constants,

$$\bar{p} = i_m^2 R \langle \sin^2 \omega t \rangle \qquad [10.5(b)]$$

Using the trigonometric identity, $\sin^2 \omega t = 1/2\ (1 - \cos 2\omega t)$, we have $\langle \sin^2 \omega t \rangle = (1/2)(1 - \langle \cos 2\omega t \rangle)$ and since $\langle \cos 2\omega t \rangle = 0$**, we have,

$$\langle \sin^2 \omega t \rangle = \frac{1}{2}$$

Thus,

$$\bar{p} = \frac{1}{2} i_m^2 R \qquad [10.5(c)]$$

To express ac power in the same form as dc power ($P = i^2R$), a special value of current is defined and used. It is called, *root mean square* (rms) or *effective current* (Fig. 10.3) and is denoted by I_{rms} or I.

*The average value of a function $F(t)$ over a period T is given by $F(t) = \frac{1}{T}\int_0^T F(t)dt$

**$\langle \cos 2\omega t \rangle = \frac{1}{T}\int_0^T \cos 2\omega t\, dt = \frac{1}{T}\left[\frac{\sin 2\omega t}{2\omega}\right]_0^T = \frac{1}{2\omega T}[\sin 2\omega T - 0] = 0$

It is defined by

$$I = \sqrt{\overline{i^2}} = \sqrt{\frac{1}{2} r_m^2} = \frac{i_m}{\sqrt{2}}$$

$$= 0.707\ i_m \qquad (10.6)$$

In terms of I, the average power, denoted by P is

$$P = \bar{p} = \frac{1}{2} i_m^2 R = I^2 R \qquad (10.7)$$

Similarly, we define the *rms voltage* or *effective voltage* by

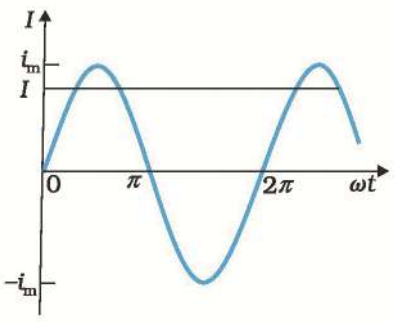


Fig. 10.3 The rms current I is related to the peak current i_m by $I = i_m / 2 = 0.707\ i$.

$$V = \frac{v_m}{\sqrt{2}} = 0.707\, v_m \qquad (10.8)$$

From Eq. (10.3), we have

$$v_m = i_m R$$

or, $$\frac{v_m}{\sqrt{2}} = \frac{i_m}{\sqrt{2}} R$$

or, $$V = IR \qquad (10.9)$$

Equation (10.9) gives the relation between ac current and ac voltage and is similar to that in the dc case. This shows the advantage of introducing the concept of rms values. In terms of rms values, the equation for power [Eq. (10.7)] and relation between current and voltage in ac circuits are essentially the same as those for the dc case.

It is customary to measure and specify rms values for ac quantities. For example, the household line voltage of 220 V is an rms value with a peak voltage of

$$v_m = \sqrt{2}\, V = (1.414)(220\ \text{V}) = 311\ \text{V}$$

In fact, the I or rms current is the equivalent dc current that would produce the same average power loss as the alternating current. Equation (10.7) can also be written as

$$P = V^2 / R = I\, V \qquad (\text{since } V = I\, R)$$

10.3 REPRESENTATION OF AC CURRENT AND VOLTAGE BY ROTATING VECTORS — PHASORS

In the previous section, we learnt that the current through a resistor is in phase with the ac voltage. But this is not so in the case of an inductor, a capacitor or a combination of these circuit elements. In order to show phase relationship between voltage and current in an ac circuit, we use the notion of *phasors*. The analysis of an ac circuit is facilitated by the use of a phasor diagram. A phasor* is a vector which rotates about the origin with angular speed ω, as shown in Fig. 10.4. The vertical components of phasors **V** and **I** represent the sinusoidally varying quantities v and i. The magnitudes of phasors **V** and **I** represent the amplitudes or the peak values v_m and i_m of these oscillating quantities. Figure 10.4(a) shows the voltage and current phasors and their relationship at time t_1 for the case of an ac source connected to a resistor i.e., corresponding to the circuit shown in Fig. 10.1. The projection of voltage and current phasors on vertical axis, i.e., $v_m \sin \omega t$ and $i_m \sin \omega t$, respectively represent the value of voltage and current at that instant. As they rotate with frequency ω, curves in Fig. 10.4(b) are generated.

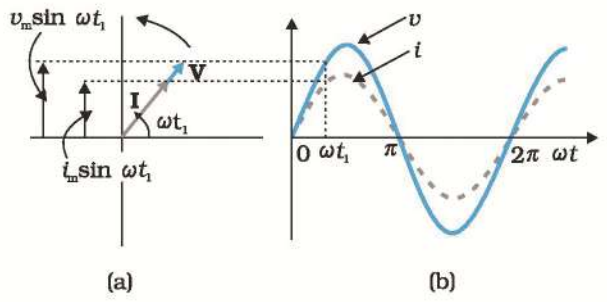


Fig. 10.4 (a) A phasor diagram for the circuit in Fig 10.1. (b) Graph of v and i versus ωt.

From Fig. 10.4(a) we see that phasors **V** and **I** for the case of a resistor are in the same direction. This is so for all times. This means that the phase angle between the voltage and the current is zero.

10.4 AC VOLTAGE APPLIED TO AN INDUCTOR

Figure 10.5 shows an ac source connected to an inductor. Usually, inductors have appreciable resistance in their windings, but we shall assume that this inductor has negligible resistance. Thus, the circuit is a purely inductive ac circuit. Let the voltage across the source be $v = v_m \sin \omega t$. Using the Kirchhoff's loop rule, $\sum \varepsilon(t) = 0$, and since there is no resistor in the circuit,

Fig. 10.5 An ac source connected to an inductor.

$$v - L\frac{di}{dt} = 0 \qquad (10.10)$$

where the second term is the self-induced Faraday emf in the inductor; and L is the self-inductance of the inductor. The negative sign follows from Lenz's law (Chapter 9). Combining Eqs. (10.1) and (10.10), we have

$$\frac{di}{dt} = \frac{v}{L} = \frac{v_m}{L} \sin \omega t \qquad (10.11)$$

Equation (10.11) implies that the equation for $i(t)$, the current as a function of time, must be such that its slope di/dt is a sinusoidally varying quantity, with the same phase as the source voltage and an amplitude given by v_m/L. To obtain the current, we integrate di/dt with respect to time:

$$\int \frac{di}{dt} dt = \frac{v_m}{L} \int \sin(\omega t) dt$$

and get,

$$i = -\frac{v_m}{\omega L} \cos(\omega t) + \text{constant}$$

The integration constant has the dimension of current and is time- independent. Since the source has an emf which oscillates symmetrically about zero, the current it sustains also oscillates symmetrically about zero, so that no constant or time-independent component of the current exists. Therefore, the integration constant is zero.

Using

$-\cos(\omega t) = \sin\left(\omega t - \frac{\pi}{2}\right)$, we have

$$i = i_m \sin\left(\omega t - \frac{\pi}{2}\right) \qquad (10.12)$$

where $i_m = \frac{v_m}{\omega L}$ is the amplitude of the current. The quantity ωL is analogous to the resistance and is called *inductive reactance*, denoted by X_L:

$$X_L = \omega L \tag{10.13}$$

The amplitude of the current is, then

$$i_m = \frac{v_m}{X_L} \tag{10.14}$$

The dimension of inductive reactance is the same as that of resistance and its SI unit is ohm (Ω). The inductive reactance limits the current in a purely inductive circuit in the same way as the resistance limits the current in a purely resistive circuit. The inductive reactance is directly proportional to the inductance and to the frequency of the current.

A comparison of Eqs. (10.1) and (10.12) for the source voltage and the current in an inductor shows that the current lags the voltage by π/2 or one-quarter (1/4) cycle. Figure 10.6 (a) shows the voltage and the current phasors in the present case at instant *t*1. The current phasor **I** is π/2 behind the voltage phasor **V**. When rotated with frequency ω counter-clockwise, they generate the voltage and current given by Eqs. (10.1) and (10.12), respectively and as shown in Fig. 10.6(b).

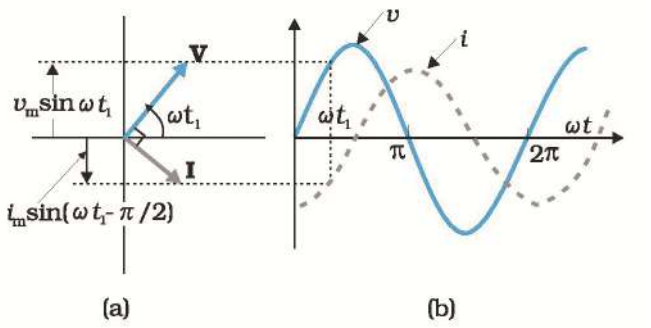


Fig.10.6 (a) A Phasor diagram for the circuit in Fig. 10.5 (b) Graph of *v* and *i* versus ω*t*.

We see that the current reaches its maximum value later than the voltage by one-fourth of a period $\left[\frac{T}{4} = \frac{\pi/2}{\omega}\right]$. You have seen that an inductor has reactance that limits current similar to resistance in a dc circuit. Does it also consume power like a resistance? Let us try to find out.

The instantaneous power supplied to the inductor is

$$p_L = iv = i_m \sin\left(\omega t - \frac{\pi}{2}\right) \times v_m \sin(\omega t)$$

$$= -i_m v_m \cos(\omega t) \sin(\omega t)$$

$$= -\frac{i_m v_m}{2} \sin(2\omega t)$$

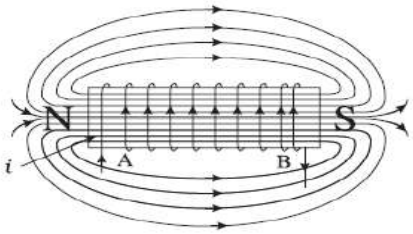
So, the average power over a complete cycle is

$$p_L = -\left\langle -\frac{i_m v_m}{2}\sin(2\omega t)\right\rangle = -\frac{i_m v_m}{2}\,[\sin(2\,\omega t)] = 0,$$

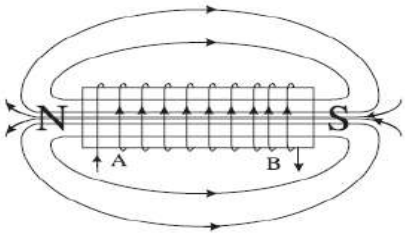
since the average of sin $(2\omega t)$ over a complete cycle is zero. Thus, the *average power supplied to an inductor over one complete cycle is zero.*

Figure10.7explains it.

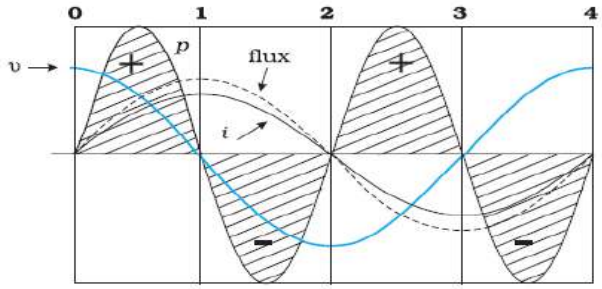
.



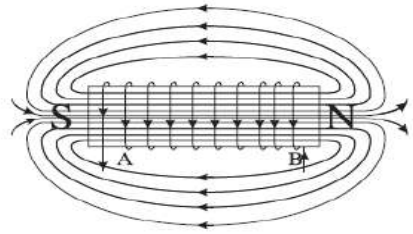
0-1 Current i through the coil entering at A increase from zero to a maximum value. Flux lines are set up i.e., the core gets magnetised. With the polarity shown voltage and current are both positive. So their product p is positive. ENERGY IS ABSORBED FROM THE SOURCE.



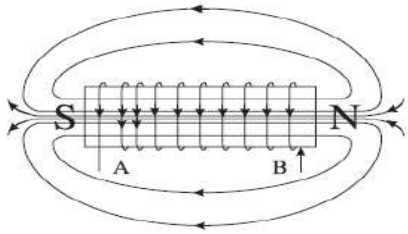
1-2 Current in the coil is still positive but is decreasing. The core gets demagnetised and the net flux becomes zero at the end of a half cycle. The voltage v is negative (since di/dt is negative). The product of voltage and current is negative, and ENERGY IS BEING RETURNED TO SOURCE.



One complete cycle of voltage/current. Note that the current lags the voltage.



2-3 Current i becomes negative i.e., it enters at B and comes out of A. Since the direction of current has changed, the polarity of the magnet changes. The current and voltage are both negative. So their product p is positive. ENERGY IS ABSORBED.



3-4 Current i decreases and reaches its zero value at 4 when core is demagnetised and flux is zero. The voltage is positive but the current is negative. The power is, therefore, negative. ENERGY ABSORBED DURING THE 1/4 CYCLE 2-3 IS RETURNED TO THE SOURCE.

Fig. 10.7 Magnetisation and demagnetisation of an inductor.

10.5 AC VOLTAGE APPLIED TO A CAPACITOR

Figure 10.8 shows an ac source ε generating ac voltage $v = v_m \sin \omega t$ connected to a capacitor only, a purely capacitive ac circuit.

When a capacitor is connected to a voltage source in a dc circuit, current will flow for the short time required to charge the capacitor. As charge accumulates on the capacitor plates, the voltage across them increases, opposing the current. That is, a capacitor in a dc circuit will limit or oppose the current as it charges. When the capacitor is fully charged, the current in the circuit falls to zero.

When the capacitor is connected to an ac source, as in Fig. 10.8, it limits or regulates the current, but does not completely prevent the flow of charge. The capacitor is alternately charged and discharged as the current reverses each half cycle. Let q be the charge on the capacitor at any time t. The instantaneous voltage v across the capacitor is

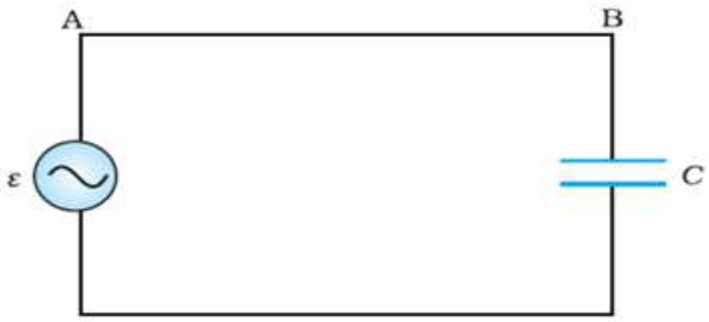


Fig 10.8 An ac source connected to a capacitor

$$v = \frac{q}{C} \qquad (10.15)$$

From the Kirchhoff's loop rule, the voltage across the source and the capacitor are equal, $v_m \sin \omega t = \frac{q}{C}$

To find the current, we use the relation $i = \frac{dq}{dt}$

$$i = \frac{d}{dt}(v_m C \sin \omega t) = \omega C v_m \cos(\omega t)$$

Using the relation, $\cos(\omega t) = \sin\left(\omega t + \frac{\pi}{2}\right)$, we have

$$i = i_m \sin\left(\omega t + \frac{\pi}{2}\right) \qquad (10.16)$$

where the amplitude of the oscillating current is $i_m = \omega C v_m$. We can rewrite it as

$$i_m = \frac{v_m}{(1/\omega C)}$$

Comparing it to $i_m = v_m/R$ for a purely resistive circuit, we find that $(1/\omega C)$ plays the role of resistance. It is called *capacitive reactance* and is denoted by X_c,

$$X_c = 1/\omega C \qquad (10.17)$$

so that the amplitude of the current is

$$i_m = \frac{v_m}{X_C} \qquad (10.18)$$

The dimension of capacitive reactance is the same as that of resistance and its SI unit is ohm (Ω). The capacitive reactance limits the amplitude of the current in a purely capacitive circuit in the same way as the resistance limits the current in a purely resistive circuit. But it is inversely proportional to the frequency and the capacitance.

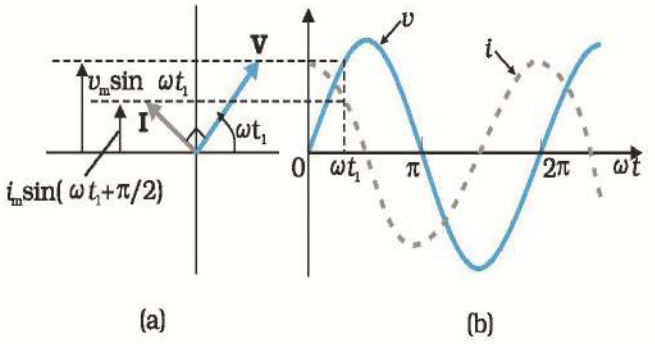


Fig. 10.9 (a) A Phasor diagram for the circuit in Fig. 10.8. (b) Graph of *v* and *i* versus ω*t*.

A comparison of Eq. (10.16) with the equation of source voltage, Eq. (10.1) shows that the current is π/2 ahead of voltage. Figure 10.9(a) shows the phasor diagram at an instant t_1. Here the current phasor **I** is π/2 ahead of the voltage phasor **V** as they rotate counter clockwise. Figure 10.9(b) shows the variation of voltage and current with time. We see that the current reaches its maximum value earlier than the voltage by one-fourth of a period.

The instantaneous power supplied to the capacitor is

$$p_c = i\,v = i_m \cos(\omega t)\, v_m \sin(\omega t)$$

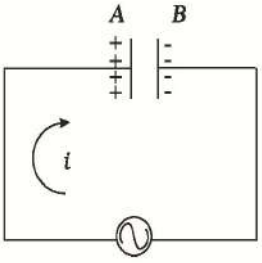
$$= i_m v_m \cos(\omega t) \sin(\omega t)$$

$$= \frac{i_m v_m}{2} \sin(2\omega t) \qquad (10.19)$$

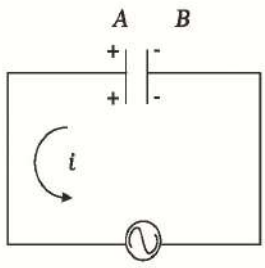
So, as in the case of an inductor, the average power

$$p_c = \left\langle \frac{i_m v_m}{2} \sin(2\omega t) \right\rangle = \frac{i_m v_m}{2} \langle \sin(2\omega t) \rangle = 0$$

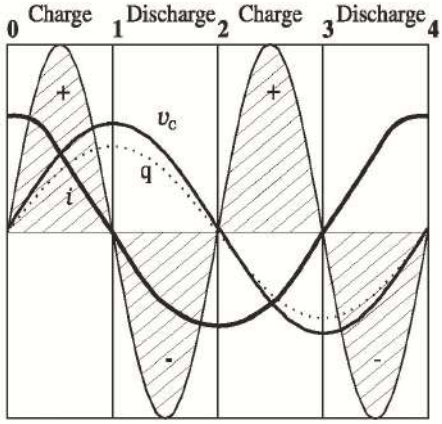
since <sin (2ω*t*)> = 0 over a complete cycle. Figure 10.10 explains it in detail. Thus, we see that in the case of an inductor, the current lags the voltage by π/2 and in the case of a capacitor, the current leads the voltage by π/2.



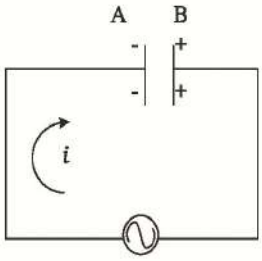
0-1 The current i flows as shown and from the maximum at 0, reaches a zero value at 1. The plate A is charged to positive polarity while negative charge q builds up in B reaching a maximum at 1 until the current becomes zero. The voltage $v_c = q/C$ is in phase with q and reaches maximum value at *1*. Current and voltage are both positive. So $p = v_c i$ is positive. ENERGY IS ABSORBED FROM THE SOURCE DURING THIS QUARTER CYCLE AS THE CAPACITOR IS CHARGED.



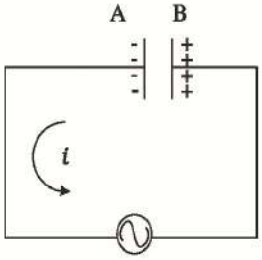
1-2 The current i reverses its direction. The accumulated charge is depleted i.e., the capacitor is discharged during this quarter cycle.The voltage gets reduced but is still positive. The current is negative. Their product, the power is negative.
THE ENERGY ABSORBED DURING THE 1/4 CYCLE **0-1** IS RETURNED DURING THIS QUARTER.



One complete cycle of voltage/current. Note that the current leads the voltage.



2-3 As i continues to flow from A to B, the capacitor is charged to reversed polarity i.e., the plate B acquires positive and A acquires negative charge. Both the current and the voltage are negative. Their product p is positive. The capacitor ABSORBS ENERGY during this 1/4 cycle.



3-4 The current i reverses its direction at **3** and flows from B to A. The accumulated charge is depleted and the magnitude of the voltage v_c is reduced. v_c becomes zero at **4** when the capacitor is fully discharged. The power is negative.ENERGY ABSORBED DURING **2-3** IS RETURNED TO THE SOURCE. NET ENERGY ABSORBED IS ZERO.

Fig. 10.10 Charging and discharging of a capacitor.

10.6 AC VOLTAGE APPLIED TO A SERIES LCR CIRCUIT

Figure 10.11 shows a series LCR circuit connected to an ac source ε. As usual, we take the voltage of the source to be $v = v_m \sin \omega t$.

If q is the charge on the capacitor and i the current, at time t, we have, from Kirchhoff's loop rule :

$$L\frac{di}{dt} + i R + \frac{q}{C} = v \qquad (10.20)$$

We want to determine the instantaneous current i and its phase relationship to the applied alternating voltage v. We shall solve this problem by two methods. First, we use the technique of phasors and in the second method, we solve Eq. (10.20) analytically to obtain the time–dependence of i.

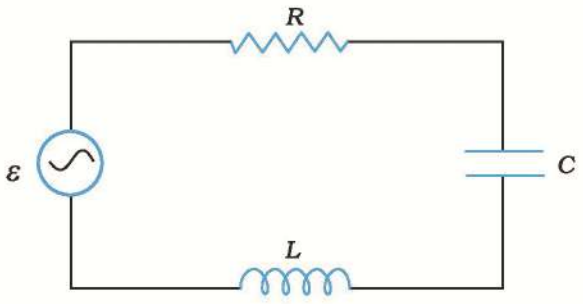


Fig. 10.11 A series *LCR* circuit connected to an ac source.

10.6.1 Phasor-diagram solution

From the circuit shown in Fig. 10.12, we see that the resistor, inductor and capacitor are in series. Therefore, the ac current in each element is the same at any time, having the same amplitude and phase. Let it be

$$i = i_m \sin(\omega t + \phi) \qquad (10.21)$$

where ϕ is the phase difference between the voltage across the source and the current in the circuit. On the basis of what we have learnt in the previous sections, we shall construct a phasor diagram for the present case.

Let **I** be the phasor representing the current in the circuit as given by Eq. (10.21). Further, let $\mathbf{V_L}$, $\mathbf{V_R}$, $\mathbf{V_C}$, and **V** represent the voltage across the inductor, resistor, capacitor and the source, respectively. From previous section, we know that $\mathbf{V_R}$is parallel to **I**, $\mathbf{V_C}$is $\pi/2$ behind **I** and $\mathbf{V_L}$is $\pi/2$ ahead of **I**. $\mathbf{V_L}$, $\mathbf{V_R}$, $\mathbf{V_C}$and **I** are shown in Fig. 10.12(a) with apppropriate phase- relations.

The length of these phasors or the amplitude of $\mathbf{V_R}$, $\mathbf{V_C}$and $\mathbf{V_L}$are :

$$v_{Rm} = i_m R,\ v_{Cm} = i_m X_C,\ v_{Lm} = i_m X_L \qquad (10.22)$$

The voltage Equation (10.20) for the circuit can be written as

$$v_L + v_R + v_C = v \qquad (10.23)$$

The phasor relation whose vertical component gives the above equation is

$$\mathbf{V_L} + \mathbf{V_R} + \mathbf{V_C} = \mathbf{V} \qquad (10.24)$$

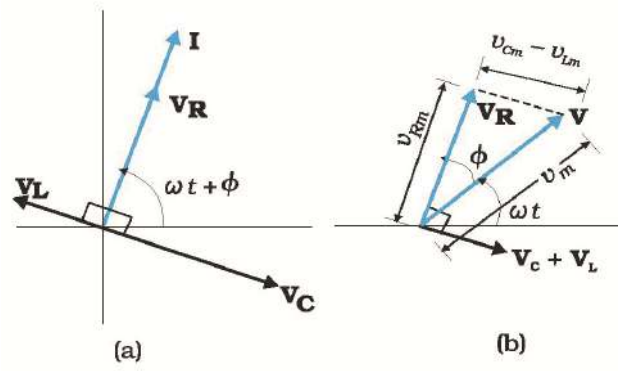


Fig.10.12 (a) Relation between the phasors $\mathbf{V_L}$, $\mathbf{V_R}$, $\mathbf{V_C}$, and *I*,

(b) Relation between the phasors $\mathbf{V_L}$, $\mathbf{V_R}$, and ($\mathbf{V_L}$ + $\mathbf{V_C}$) for the circuit in Fig. 10.11.

This relation is represented in Fig. 10.12(b). Since $\mathbf{V_C}$and $\mathbf{V_L}$are always along the same line and in opposite directions, they can be combined into a single phasor ($\mathbf{V_C}$+ $\mathbf{V_L}$) which has a magnitude $|v_{Cm} - v_{Lm}|$. Since **V** is represented as the hypotenuse of a right-triangle whose sides are $\mathbf{V_R}$and ($\mathbf{V_C}$+ $\mathbf{V_L}$), the pythagorean theorem gives :

$$v_m^2 = v_{Rm}^2 + (v_{Cm} - v_{Lm})^2$$

Substituting the values of v_{Rm}, v_{Cm}, and $v_{Lm}$from Eq. (10.22) into the above equation, we have

$$v_m^2 = (i_m R)^2 + (i_m X_C - i_m X_L)^2$$
$$= i_m^2 [R^2 + (X_C - X_L)^2]$$

$$\text{or, } i_m = \frac{v_m}{\sqrt{R^2 + (X_C - X_L)^2}} \qquad [10.25(a)]$$

By analogy to the resistance in a circuit, we introduce the *impedance Z* in an ac circuit :

$$i_m = \frac{v_m}{Z} \qquad [10.25(b)]$$

$$\text{where } Z = \sqrt{R^2 + (X_C - X_L)^2} \qquad (10.26)$$

Since phasor **I** is always parallel to phasor $\mathbf{V_R}$, the phase angle ϕ is the angle between $\mathbf{V_R}$and **V** and can be determined from Fig. 10.13 :

$$\tan\phi = \frac{v_{Cm} - v_{Lm}}{v_{Rm}}$$

Using Eq. (10.22), we have

$$\tan\phi = \frac{X_C - X_L}{R} \qquad (10.27)$$

Equations (10.26) and (10.27) are graphically shown in Fig. (10.13). This is called *Impedance diagram* which is a right-triangle with Z as its hypotenuse.

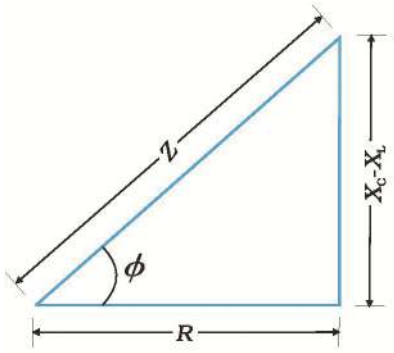


Fig. 10.13 Impedance diagram.

Equation 10.25(a) gives the amplitude of the current and Eq. (10.27) gives the phase angle. With these, Eq. (10.21) is completely specified.

If $X_C > X_L$, ϕ is positive and the circuit is predominantly capacitive. Consequently, the current in the circuit leads the source voltage. If $X_C < X_L$, ϕ is negative and the circuit is predominantly inductive. Consequently, the current in the circuit lags the source voltage.

Figure 10.15 shows the phasor diagram and variation of v and i with ωt for the case $X_C > X_L$.

Thus, we have obtained the amplitude and phase of current for an *LCR* series circuit using the technique of phasors. But this method of analysing ac circuits suffers from certain disadvantages. First, the phasor diagram say nothing about the initial condition. One can take any arbitrary value of t (say, $t1$, as done throughout this chapter) and draw different phasors which show the relative angle between different phasors. The solution so obtained is called the *steady-state solution*. This is not a general solution. Additionally, we do have a *transient solution* which exists even for $v = 0$. The general solution is the sum of the transient solution and the steady-state solution. After a sufficiently long time, the effects of the transient solution die out and the behaviour of the circuit is described by the steady-state solution.

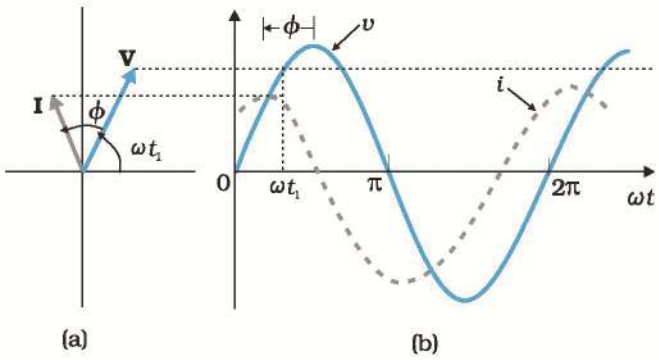


Fig.10.14 (a) Phasor diagram of **V** and **I**. (b) Graphs of v and i versus ωt for a series *LCR* circuit where $X_C > X_L$.

10.6.2 Analytical solution

The voltage equation for the circuit is

$$L\frac{di}{dt} + R\,i + \frac{q}{C} = v$$

$$= v_m \sin \omega t$$

We know that $i = dq/dt$. Therefore, $di/dt = d^2q/dt^2$. Thus, in terms of q, the voltage equation becomes

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{q}{C} = v_m \sin \omega t \qquad (10.28)$$

This is like the equation for a forced, damped oscillator, [in Class XI Physics Textbook]. Let us assume a solution

$$q = q_m \sin(\omega t + \theta) \qquad [10.29(a)]$$

so that $\frac{dq}{dt} = q_m \omega^2 \cos(\omega t + \theta)$ [10.29(b)]

and $\frac{d^2q}{dt^2} = -q_m \omega^2 \sin(\omega t + \theta)$ [10.29(c)]

Substituting these values in Eq. (10.28), we get

$$q_m \omega[R \cos(\omega t + \theta) + (X_C - X_L) \sin(\omega t + \theta)] = v_m \sin \omega t \qquad (10.30)$$

where we have used the relation $X_C = 1/\omega C$, $X_L = \omega L$. Multiplying and dividing Eq. (10.30) by $Z = \sqrt{R^2 + (X_C - X_L)^2}$, we have

$$q_m \omega Z\left[\frac{R}{Z}\cos(\omega t + \theta) + \frac{(X_C - X_L)}{Z}\sin(\omega t + \theta)\right] = v_m \sin \omega t \qquad (10.31)$$

Now, let $\frac{R}{Z} = \cos \phi$

and $\frac{(X_C - X_L)}{Z} = \sin \phi$

so that $\phi = \tan^{-1}\frac{X_C - X_L}{R}$ (10.32)

Substituting this in Eq. (10.31) and simplifying, we get :

$$q_m \omega Z \cos(\omega t + \theta - \phi) = v_m \sin \omega t \qquad (10.33)$$

Comparing the two sides of this equation, we see that

$$v_m = q_m \omega Z = i_m Z$$

where

$$i_m = q_m \omega \qquad [10.33(a)]$$

and $\theta - \phi = -\frac{\pi}{2}$ or $\theta = -\frac{\pi}{2} + \phi$ [10.33(b)]

Therefore, the current in the circuit is

$$i = \frac{dq}{dt} = q_m \omega \cos(\omega t + \theta)$$

$$= i_m \cos(\omega t + \theta)$$

or $i = i_m \sin(\omega t + \phi)$ (10.34)

where $i_m = \frac{v_m}{Z} = \frac{v_m}{\sqrt{R^2 + (X_C - X_L)^2}}$ [10.34(a)]

and $\phi = \tan^{-1}\frac{X_C - X_L}{R}$

Thus, the analytical solution for the amplitude and phase of the current in the circuit agrees with that obtained by the technique of phasors.

10.6.3 Resonance

An interesting characteristic of the series *RLC* circuit is the phenomenon of resonance. The phenomenon of resonance is common among systems that have a tendency to oscillate at a particular frequency. This frequency is called the system's *natural frequency*. If such a system is driven by an energy source at a frequency that is near the natural frequency, the amplitude of oscillation is found to be large. A familiar example of this phenomenon is a child on a swing. The swing has a natural frequency for swinging back and forth like a pendulum. If the child pulls on the rope at regular intervals and the frequency of the pulls is almost the same as the frequency of swinging, the amplitude of the swinging will be large (Chapter 14, Class XI).

For an *RLC* circuit driven with voltage of amplitude *vm* and frequency ω, we found that the current amplitude is given by

$$i_m = \frac{v_m}{Z} = \frac{v_m}{\sqrt{R^2 + (X_C - X_L)^2}}$$

with $X_C = 1/\omega C$ and $X_L = \omega L$. So if ω is varied, then at a particular frequency ω_0, $X_C = X_L$, and the impedance is minimum $\left(Z = \sqrt{R^2 + 0^2} = R\right)$. This frequency is called the *resonant frequency* :

$$X_C = X_L \text{ or } \frac{1}{\omega_0 C} = \omega_0 L \qquad \text{or} \qquad \omega_0 = \frac{1}{\sqrt{LC}} \tag{10.35}$$

At resonant frequency, the current amplitude is maximum; $i_m = v_m/R$.

Figure 10.15 shows the variation of i_m with ω in a *RLC* series circuit with L = 1.00 mH, C = 1.00 nF for two values of R: (i) R = 100 Ω and (ii) R = 200 Ω. For the source applied v_m = 100 V. ω_0 for this case is $\frac{1}{\sqrt{LC}} = 1.00 \times 10^6$ rad/s.

We see that the current amplitude is maximum at the resonant frequency. Since $i_m = v_m / R$ at resonance, the current amplitude for case (i) is twice to that for case (ii).

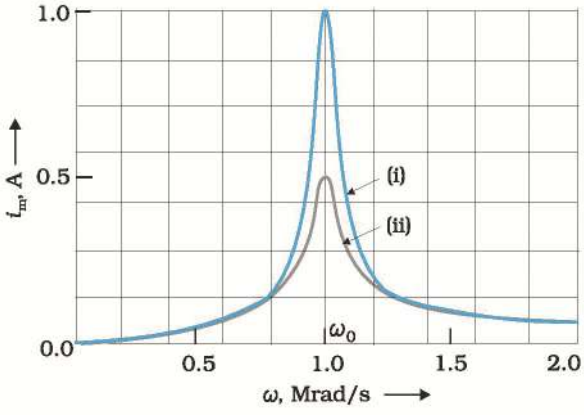


Fig. 10.15 Variation of i_m with ω for two cases : (i) R = 100 Ω, (ii) R = 200 Ω, L = 1.00 mH.

Resonant circuits have a variety of applications, for example, in the tuning mechanism of a radio or a TV set. The antenna of a radio accepts signals from many broadcasting stations. The signals picked up in the antenna acts as a source in the tuning circuit of the radio, so the circuit can be driven at many frequencies. But to hear one particular radio station, we tune the radio. In tuning, we vary the capacitance of a

capacitor in the tuning circuit such that the resonant frequency of the circuit becomes nearly equal to the frequency of the radio signal received. When this happens, the amplitude of the current with the frequency of the signal of the particular radio station in the circuit is maximum.

It is important to note that resonance phenomenon is exhibited by a circuit only if both L and C are present in the circuit. Only then do the voltages across L and C cancel each other (both being out of phase) and the current amplitude is vm/R, the total source voltage appearing across R. This means that we cannot have resonance in a RL or RC circuit.

10.7 POWER IN AC CIRCUIT: THE POWER FACTOR

We have seen that a voltage $v = v_m \sin \omega t$ applied to a series *RLC* circuit drives a current in the circuit given by $i = i_m \sin(\omega t + \phi)$ where

$$i_m = \frac{v_m}{Z} \text{ and } \phi = \tan^{-1}\left(\frac{X_C - X_L}{R}\right)$$

Therefore, the instantaneous power *p* supplied by the source is

$$p = v\, i = (v_m \sin \omega t) \times [i_m \sin(\omega t + \phi)]$$

$$= \frac{v_m i_m}{2}[\cos \phi - \cos(2\omega t + \phi)] \qquad (10.36)$$

The average power over a cycle is given by the average of the two terms in R.H.S. of Eq. (10.36). It is only the second term which is time-dependent. Its average is zero (the positive half of the cosine cancels the negative half). Therefore,

$$P = \frac{v_m i_m}{2} \cos \phi = \frac{v_m}{\sqrt{2}} \frac{i_m}{\sqrt{2}} \cos \phi$$

$$= V I \cos \phi \qquad [10.37(a)]$$

This can also be written as,

$$P = I^2 Z \cos \phi \qquad [10.37(b)]$$

So, the average power dissipated depends not only on the voltage and current but also on the cosine of the phase angle ϕ between them. The quantity $\cos\phi$ is called the *power factor*. Let us discuss the following cases:

Case (i) *Resistive circuit*: If the circuit contains only pure *R*, it is called *resistive*. In that case $\phi = 0$, $\cos \phi = 1$. There is maximum power dissipation.

Case (ii) *Purely inductive or capacitive circuit*: If the circuit contains only an inductor or capacitor, we know that the phase difference between voltage and current is $\pi/2$. Therefore, $\cos \phi = 0$, and no power is dissipated even though a current is flowing in the circuit. This current is sometimes referred to as *wattless current*.

Case (iii) *LCR series circuit*: In an *LCR* series circuit, power dissipated isgiven by Eq. (10.37) where $\phi = \tan^{-1}(X_C - X_L)/R$. So, ϕ may be non-zero in a *RL* or *RC* or *RCL* circuit. Even in such cases, power is dissipated only in the resistor.

Case (iv) *Power dissipated at resonance in LCR circuit*: At resonance $X_C - X_L = 0$, and $\phi = 0$. Therefore, $\cos \phi = 1$ and $P = I^2Z = I^2R$. That is, maximum power is dissipated in a circuit (through *R*) at resonance.

10.9 TRANSFORMERS

For many purposes, it is necessary to change (or transform) an alternating voltage from one to another of greater or smaller value. This is done with a device called *transformer* using the principle of mutual induction.

A transformer consists of two sets of coils, insulated from each other. They are wound on a soft-iron core, either one on top of the other as in Fig. 10.16(a) or on separate limbs of the core as in Fig. 10.16(b). One of the coils called the *primary coil* has N_p turns. The other coil is called the *secondary coil*; it has N_s turns. Often the primary coil is the input coil and the secondary coil is the output coil of the transformer.

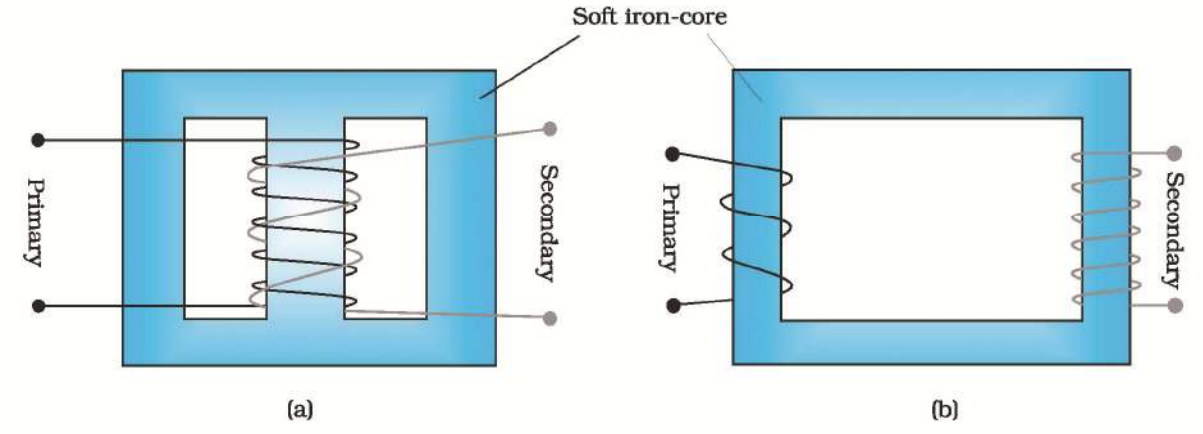


Fig. 10.16 Two arrangements for winding of primary and secondary coil in a transformer :

(a) two coils on top of each other, (b) two coils on separate limbs of the core.

When an alternating voltage is applied to the primary, the resulting current produces an alternating magnetic flux which links the secondary and induces an emf in it. The value of this emf depends on the number of turns in the secondary. We consider an ideal transformer in which the primary has negligible resistance and all the flux in the core links both primary and secondary windings. Let ϕ be the flux in each turn in the core at time t due to current in the primary when a voltage v_p is applied to it.

Then the induced emf or voltage ε_s, in the secondary with N_s turns is

$$\varepsilon_s = -N_s \frac{d\phi}{dt} \tag{10.38}$$

The alternating flux ϕ also induces an emf, called back emf in the primary. This is

$$\varepsilon_p = -N_p \frac{d\phi}{dt} \tag{10.39}$$

But $\varepsilon_p = v_p$. If this were not so, the primary current would be infinite since the primary has zero resistance (as assumed). If the secondary is an open circuit or the current taken from it is small, then to a good approximation

$$\varepsilon_s = v_s$$

where v_s is the voltage across the secondary. Therefore, Eqs. (10.38) and (10.39) can be written as

$$v_s = -N_s \frac{d\phi}{dt} \qquad [10.38(a)]$$

$$v_p = -N_p \frac{d\phi}{dt} \qquad [10.39(a)]$$

From Eqs. [10.38(a)] and [10.39(a)], we have

$$\frac{v_s}{v_p} = \frac{N_s}{N_p} \qquad [10.40]$$

Note that the above relation has been obtained using three assumptions: (i) the primary resistance and current are small; (ii) the same flux links both the primary and the secondary as very little flux escapes from the core, and (iii) the secondary current is small.

If the transformer is assumed to be 100% efficient (no energy losses), the power input is equal to the power output, and since $p = i\,v$,

$$i_p v_p = i_s v_s \qquad (10.41)$$

Although some energy is always lost, this is a good approximation, since a well designed transformer may have an efficiency of more than 95%. Combining Eqs. (10.39) and (10.41), we have

$$\frac{i_p}{i_s} = \frac{v_s}{v_p} = \frac{N_s}{N_p} \qquad (10.42)$$

Since i and v both oscillate with the same frequency as the ac source, Eq. (10.42) also gives the ratio of the amplitudes or rms values of corresponding quantities.

Now, we can see how a transformer affects the voltage and current.
We have :

$$V_s = \left(\frac{N_s}{N_p}\right) V_p \quad \text{and} \quad I_s = \left(\frac{N_p}{N_s}\right) I_p \qquad (10.43)$$

That is, if the secondary coil has a greater number of turns than the primary ($N_s > N_p$), the voltage is stepped up ($V_s > V_p$). This type of arrangement is called a *step-up transformer*. However, in this arrangement, there is less current in the secondary than in the primary ($N_p/N_s < 1$ and $I_s < I_p$). For example, if the primary coil of a transformer has 100 turns and the secondary has 200 turns, $N_s/N_p = 2$ and $N_p/N_s = 1/2$. Thus, a 220V input at 10A will step-up to 440 V output at 5.0 A.

If the secondary coil has less turns than the primary ($N_s < N_p$), we have a *step-down transformer*. In this case, $V_s < V_p$ and $I_s > I_p$. That is, the voltage is stepped down, or reduced, and the current is increased.

The equations obtained above apply to ideal transformers (without any energy losses). But in actual transformers, small energy losses do occur due to the following reasons:

(i) *Flux Leakage*: There is always some flux leakage; that is, not all of the flux due to primary passes through the secondary due to poor design of the core or the air gaps in

the core. It can be reduced by winding the primary and secondary coils one over the other.

(ii) *Resistance of the windings*: The wire used for the windings has some resistance and so, energy is lost due to heat produced in the wire (I^2R). In high current, low voltage windings, these are minimised by using thick wire.

(iii) *Eddy currents*: The alternating magnetic flux induces eddy currents in the iron core and causes heating. The effect is reduced by using a laminated core.

(iv) *Hysteresis*: The magnetisation of the core is repeatedly reversed by the alternating magnetic field. The resulting expenditure of energy in the core appears as heat and is kept to a minimum by using a magnetic material which has a low hysteresis loss.

The large scale transmission and distribution of electrical energy over long distances is done with the use of transformers. The voltage output of the generator is stepped-up (so that current is reduced and consequently, the I^2R loss is cut down). It is then transmitted over long distances to an area sub-station near the consumers. There the voltage is stepped down. It is further stepped down at distributing sub-stations and utility poles before a power supply of 240 V reaches our homes.

SUMMARY

1. An alternating voltage $v = v_m \sin \omega t$ applied to a resistor R drives a current $i = i_m \sin \omega t$ in the resistor, $i_m = \frac{v_m}{R}$. The current is in phase with the applied voltage.

2. For an alternating current $i = i_m \sin \omega t$ passing through a resistor R, the average power loss P (averaged over a cycle) due to joule heating is $(1/2) i^2 R$. To express it in the same form as the dc power ($P = I^2R$), a special value of current is used. It is called *root mean square* (rms) *current* and is donoted by I :

   $$I = \frac{i_m}{\sqrt{2}} = 0.707\, i_m$$

   Similarly, the *rms voltage* is defined by

   $$V = \frac{v_m}{\sqrt{2}} = 0.707\, v_m$$

   We have $P = IV = I^2R$

3. An ac voltage $v = v_m \sin \omega t$ applied to a pure inductor L, drives a current in the inductor $i = i_m \sin(\omega t - \pi/2)$, where $i_m = v_m / X_L$. $X_L = \omega L$ is called *inductive reactance*. The current in the inductor lags the voltage by $\pi/2$. The average power supplied to an inductor over one complete cycle is zero.

4. An ac voltage $v = v_m \sin \omega t$ applied to a capacitor drives a current in the capacitor: $i = i_m \sin(\omega t + \pi/2)$. Here,

   $$i_m = \frac{v_m}{X_C}, \quad X_C = \frac{1}{\omega C}$$ is called *capacitive reactance*.

   The current through the capacitor is $\pi/2$ ahead of the applied voltage. As in the case of inductor, the average power supplied to a capacitor over one complete cycle is zero.

5. For a series *RLC* circuit driven by voltage $v = v_m \sin \omega t$, the current is given by $i = i_m \sin(\omega t + \phi)$

where $i_m = \dfrac{v_m}{\sqrt{R^2 + (X_C - X_L)^2}}$

and $\phi = \tan^{-1} \dfrac{X_C - X_L}{R}$

$Z = \sqrt{R^2 + (X_C - X_L)^2}$ is called the *impedance* of the circuit.

The average power loss over a complete cycle is given by

$P = V I \cos \phi$

The term $\cos \phi$ is called the *power factor*.

6. In a purely inductive or capacitive circuit, $\cos \phi = 0$ and no power is dissipated even though a current is flowing in the circuit. In such cases, current is referred to as a *wattless current*.

7. The phase relationship between current and voltage in an ac circuit can be shown conveniently by representing voltage and current by rotating vectors called *phasors*. A phasor is a vector which rotates about the origin with angular speed ω. The magnitude of a phasor represents the amplitude or peak value of the quantity (voltage or current) represented by the phasor.

   The analysis of an ac circuit is facilitated by the use of a phasor diagram.

8. A transformer consists of an iron core on which are bound a primary coil of N_p turns and a secondary coil of N_s turns. If the primary coil is connected to an ac source, the primary and secondary voltages are related by

$$V_s = \left(\frac{N_s}{N_p} \right) V_p$$

and the currents are related by

$$I_s = \left(\frac{N_p}{N_s} \right) I_p$$

If the secondary coil has a greater number of turns than the primary, the voltage is stepped-up ($V_s > V_p$). This type of arrangement is called a *step-up transformer*. If the secondary coil has turns less than the primary, we have a *step-down transformer*.

VERY SHORT ANSWER QUESTIONS (2 MARKS)

1. A transformer converts 200 V ac into 2000 V ac. Calculate the number of turns in the secondary if the primary has 10 turns.
2. What type of transformer is used in a 6V bed lamp?
3. What is the phenomenon involved in the working of a transformer?
4. What is transformer ratio?
5. Write the expression for the reactance of (i) an inductor and (ii) a capacitor
6. What is the phase difference between AC emf and current in the following: pure inductor and pure capacitor?
7. Define power factor. On which factors does power factor depend?
8. What is meant by wattles component of current?
9. When does a LCR series circuit have minimum impedance?
10. What is the phase difference between voltage and current when the power factor in LCR series circuit is unity?

SHORT ANSWER QUESTIONS (4 MARKS)

1. Obtain an expression for the current through an inductor when an AC emf is applied.
2. Obtain an expression for the current in a capacitor when an AC emf is applied.
3. State the principle on which a transformer works. Describe the working of a transformer with necessary theory.

LONG ANSWER QUESTIONS (8 MARKS)

1. Obtain an expression for impedance and current in series LCR circuit. Deduce an expression for the resonating frequency of an LCR series resonating circuit.

CHAPTER 11

ELECTROMAGNETIC WAVES

11.1 INTRODUCTION

In Chapter 7, we learnt that an electric current produces magnetic field and that two current carrying wires exert a magnetic force on each other. Further, in Chapter 9 we have seen that a magnetic field changing with time gives rise to an electric field. Is the converse also true ? Does an electric field changing with time give rise to a magnetic field? James Clerk Maxwell (1831 – 1879), argued that this was indeed the case – not only an electric field changing with time –varying electric field generates magnetic field while applying the Ampere's circuital law. He suggested the existence of an additional current, called by him, the displacement current to remove this inconsistency.

Maxwell formulated a set of equations involving electric and magnetic fields, and their sources, the charge and current densities. These equations are known as Maxwell's equations. Together with the Lorentz force formula (Chapter7), they mathematically express all the basic laws of electromagnetism.

James Clerk Maxwell (1831 – 1879) Born in Edinburgh, Scotland, was among the greatest physicists of the nineteenth century. He derived the thermal velocity distribution of molecules in a gas and was among the first to obtain reliable estimates of molecular parameters from measurable quantities like viscosity, etc. Maxwell's greatest acheivement was the unification of the laws of electricity and magnetism (discovered by Coulomb, Oersted, Ampere and Faraday) into a consistent set of equations now called Maxwell's equations. From these he arrived at the most important conclusion that light is an electromagnetic wave. Interestingly, Maxwell did not agree with the idea (strongly suggested by the Faraday's laws of electrolysis) that electricity was particulate in nature.

The most important prediction to emerge from Maxwell's equations is the existence of electromagnetic waves, which are (coupled) time-varying electric and magnetic fields that propagate in space. The speed of the waves, according to these equations, turned out to be very close to the speed of light (3×10^8 m / s) , obtained from optical measurements. This led to the remarkable conclusion that light is an electromagnetic wave. Maxwell's work thus unified the domain of electrically, magnetism and light. Hertz, in 1885 , experimentally demonstrated the existence of electromagnetic waves. Its technological use by Marconi and others led in due course to the revolution in communication that we are witnessing today.

In this chapter, we first discuss the need for displacement current and its consequences. Then we present a descriptive account of electromagnetic waves. The broad spectrum of electromagnetic waves, stretching from γ rays (wavelength – 10^{-12} m) to long radio waves (wavelength – 10^6 m) is described . How the electromagnetic waves are sent and received for communication is discussed in Chapter 16.

11.2 DISPLACEMENT CURRENT

We have seen in Chapter 7 that an electrical current produces a magnetic field around it. Maxwell showed that for logical consistency, a changing electric field must also produce a

magnetic field. This effect is of great importance because it explains the existence of radio waves, gamma rays and visible light, as well as all other forms of electromagnetic waves.

To see how a changing electric field gives rise to a magnetic field, let us consider the process of charging of a capacitor and apply Ampere's circuital law given by (Chapter 7)

$$\oint \boldsymbol{B}.\boldsymbol{d1} = \boldsymbol{\mu_0}\, \boldsymbol{i(t)} \qquad (11.1)$$

To find magnetic field at a point outside the capacitor. Figure 11.1(a) shows a parallel plate capacitor C which is a part of circuit through which a time-dependent current ***i(t)*** flows. Let us find the magnetic field at a point such as P, in a region outside the parallel plate capacitor. For this, we consider a plane circular loop of radius r whose plane is perpendicular to the direction of the current-carrying wire, and which is centred symmetrically with respect to the wire [Fig.11.1(a)] .

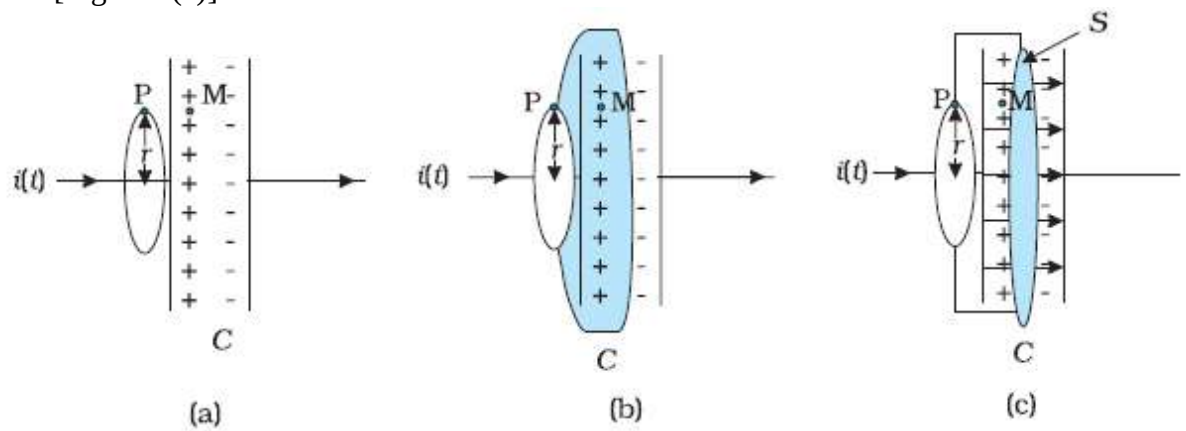


Fig. 11.1 A parallel plate capacitor *C*, as part of a circuit through which a time dependent current *i* (*t*) flows, (a) a loop of radius *r*, to determine magnetic field at a point P on the loop; (b) a pot-shaped surface passing through the interior between the capacitor plates with the loop shown in (a) as its rim; (c) a tiffin-shaped surface with the circular loop as its rim and a flat circular bottom *S* between the capacitor plates. The arrows show uniform electric field between the capacitor plates.

From symmetry, the magnetic field is directed along the circumference of the circular loop and is the same in magnitude at all points on the loop so that if B is the magnitude of the field, the left side of [Eq.11.1] is B($2\pi r$) .So we have

$$\mathbf{B}(\boldsymbol{2\pi r}) = \boldsymbol{\mu_0}\, \boldsymbol{i(t)} \qquad (11.2)$$

Now , consider a different surface , which has the same boundary. This is a pot like surface [Fig 11.1(b)] which nowhere touches the current, but has its bottom between the capacitor plates; its mouth is the circular loop mentioned above. Another such surface is shaped like a tiffin box (without the lid) [Fig 11.1(c)]. On applying Ampere's circuital law to such surfaces with the same perimeter, we find that the left hand side of Eq.[11.1] has not changed but the right hand side is zero and not $\boldsymbol{\mu_0\, i}$ **.**

Since *no current* passes through the surface of Fig 11.1(b) and (c). So we have a *contradiction***;** calculated one way, there is a magnetic field at a point P; calculated another way, the magnetic field at P is zero**.** Since the contradiction arises from our use of Ampere's circuital law, this law must be missing something. The missing term must be such that one gets the same magnetic field at point P, no matter what surface is used**.**

We can actually guess the missing term by looking carefully at [Fig,11.1(c)] . Is there anything passing through the surface S between the plates of the capacitor ? Yes, of course, the electric field ! If the plates of the capacitor have an area A, and a total charge Q, the magnitude of the electric field E between the plates is [Q/A] /ε_0 [see Eq. 5.41] . The field is perpendicular to the surface S of Fig [11.1(c)] . It has the same magnitude over the area A of the capacitor plates, and vanishes outside it. So what is the electric flux $\emptyset_E$ through the surface S ? Using Gauss's law, it is

$$\emptyset_E = |E|A = \frac{1}{\varepsilon_0}\frac{Q}{A}A = \frac{Q}{\varepsilon_0} \qquad (11.3)$$

Now if the charge Q on the capacitor plates changes with time, there is a current i = (dQ/dt), so that using Eq.(11.3) , we have

$$\frac{d\emptyset_\varepsilon}{dt} = \frac{d}{dt}\left(\frac{Q}{\varepsilon_0}\right) = \frac{1}{\varepsilon_0}\frac{dQ}{dt}$$

This implies that for consistency ,

$$\varepsilon_0\left(\frac{d\emptyset_\varepsilon}{dt}\right) = i \qquad (11.4)$$

This is the missing term in Ampere's circuital law. If we generalise this law by adding to the current carried by conductors through the surface, another term which is ε_0 times the rate of change of electric flux through the same surface, the *total* has the same value of current i for all surfaces. If this is done, there is no contradiction in the value of B obtained anywhere using the generalised Ampere's law. B at the point P is non-zero no matter which surface is used for calculating it. B at a point P outside the plates [Fig .11.1(a)] is the same as at a point M just inside, as it should be. The current carried by conductors due to flow of charges is called *conduction current.* The current, given by Eq.[11.4] , is a new term, and is due to changing electric field (or *electric displacement, an old term still used sometimes)* . It is , therefore , called *displacement, current* or Maxwell's displacement current. Fig [11.2] shows the electric and magnetic fields inside the parallel plate capacitor discussed above.

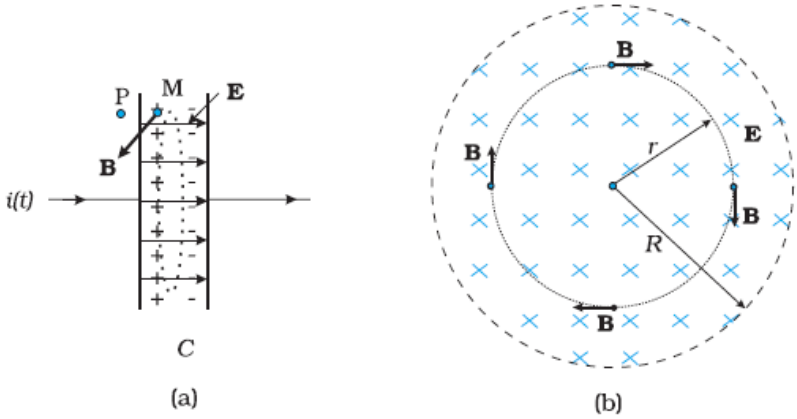


Fig. 11.2 (a) The electric and magnetic fields E and B between the capacitor plates, at the point M. (b) A cross sectional view of Fig. (a).

The generalisation made by Maxwell then is the following. The source of a magnetic field is not *just* the conduction electric current due to flowing charges, but also the time rate of charge of electric field. More precisely , the total current $\boldsymbol{i}$ is the sum of the conduction current denoted by $\boldsymbol{i_C}$ **,** and the displacement current denoted by

$$i_d (= \varepsilon_0 (d\emptyset_\varepsilon / dt)]$$

So we have

$$i = i_C + i_d = i_C + \varepsilon_0 \frac{d\emptyset_\varepsilon}{dt} \qquad (11.5)$$

In explicit terms, this means that outside the capacitor plates, we have only conduction current $i_C = i$, and no displacement current , i.e., $i_d = 0$. On the other hand, inside the capacitor, there is no conduction current , i.e., $i_C = 0$, and there is only displacement current , so that $i_d = i$.

The generalised (and correct) Ampere's circuital law has the same form as Eq.[11.1] , with one difference ; " the *total current* passing through any surface of which the closed loop is the perimeter" is the sum of the conduction current and the displacement current.

The generalised law is

$$\oint B.dI = \mu_0 i_C + \mu_0 \varepsilon_0 \frac{d\emptyset_\varepsilon}{dt} \qquad (11.6)$$

And is known as Ampere-Maxwell law.

In all respects, the displacement current has the same physical effects as the conduction current. In some cases, for example, steady electric fields in a conducting wire, the displacement current may be zero since the electric field **E** does not change with time. In other cases, for example , the charging capacitor above, both conduction and displacement currents may be present in different regions of space. In most of the cases, they both may be present in the same region of space, as there exist no perfectly conducting or perfectly insulating medium. Most interestingly, there may be large regions of space where there is no conduction current, but there is only a displacement current due to time-varying electric fields. In such a region , we expect a magnetic field, though there is no (conduction) current source nearby ! The prediction of such a displacement current can be verified experimentally. For example, a magnetic field (say at point M) between the plates of the capacitor in Fig [11.2(a)] can be measured and is seen to be the same as that just outside (at P).

The displacement current has (literally) far reaching consequences. One thing we immediately notice is that the laws of electricity and magnetism are now more symmetrical" Faraday's law of induction states that there is an induced emf *equal to the rate of change* of magnetic flux. Now, since the emf between two points 1 and 2 is the work done per unit charge in taking it from 1 to 2, the existence of an emf implies the existence of an electric field. So, we can rephrase Faraday's law of electromagnetic induction by saying that a *magnetic field,* changing with time, gives rise to an *electric field.* Then, the fact that an *electric field* changing with time gives rise to a *magnetic field,* is the symmetrical counterpart, and is a consequence of the displacement current being a source of a magnetic field. Thus, time-dependent electric and magnetic fields give rise to each other ! Faraday's law of electromagnetic induction and Ampere-Maxwell law give a quantitative expression of this

statement, with the current being the total current, as in Eq[11.5] . One very important consequence of this symmetry is the existence of electromagnetic waves, which we discuss qualitatively in the next section.

MAXWELL'S EQUATIONS

1.	$\oint \boldsymbol{E}.\boldsymbol{dA} = \boldsymbol{Q}/\boldsymbol{\varepsilon_O}$	(Gauss's Law for electricity)
2.	$\oint \boldsymbol{B}.\boldsymbol{dA} = \boldsymbol{0}$	(Gauss's Law for magnetism)
3.	$\oint \boldsymbol{E}.\boldsymbol{dI} = \frac{-d\emptyset_B}{dt}$	(Faraday's Law)
4.	$\oint \boldsymbol{B}.\boldsymbol{dI} = \boldsymbol{\mu_0 i_c} + \boldsymbol{\mu_0 \varepsilon_0} \frac{d\emptyset_E}{dt}$	(Ampere – Maxwell Law)

11.3 ELECTROMAGNETIC WAVES

11.3.1 Sources of electromagnetic waves

How are electromagnetic waves produced ? Neither stationary charges nor charges in uniform motion (steady currents) can be sources of electromagnetic waves. The former produces only electrostatic fields, while the latter produces magnetic fields that, however, do not vary with time. It is an important result of Maxwell's theory that accelerated charges radiate electromagnetic waves. The proof of this basic result is beyond the scope of this book, but we can accept it on the basis of rough , qualitative reasoning. Consider a charge oscillating with some frequency. (An oscillating charge is an example of accelerating charge.) This produces an oscillating electric field in space, which produces an oscillating magnetic field, which in turn , is a source of oscillating electric field, and so on. The oscillating electric and magnetic fields thus regenerate each other, so to speak, as the wave propagates through the space. The frequency of the electromagnetic wave naturally equals the frequency of oscillation of the charge. The energy associated with the propagating wave comes at the expense of the energy of the source – the accelerated charge.

Heinrich Rudolf Hertz (1857 – 1894) German physicist who was the first to broadcast and receive radio waves. He produced electromagnetic waves, sent them through space, and measured their wavelength and speed. He showed that the nature of their vibration, reflection and refraction was the same as that of light and heat waves, establishing their identity for the first time. He also pioneered research on discharge of electricity through gases, and discovered the photoelectric effect.

From the preceding discussion, it might appear easy to test the prediction that light is an electromagnetic wave. We might think that all we needed to do was to set up an ac circuit in which the current oscillate at the frequency of visible light, say, yellow light. But, alas, that is not possible. The frequency of yellow light is about 6 x 10^{14} Hz, while the frequency that we get even with modern electronic circuits is hardly about 10^{11} Hz. This is why the

experimental demonstration of electromagnetic wave had to come in the low frequency region (the radio wave region), as in the Hertz's experiment (1887).

Hertz's successful experimental test of Maxwell's theory created a sensation and parked off other important works in this field. Two important achievements in this onnection deserve mention. Seven years after Hertz, Jagdish Chandra Bose , working at Calcutta (now Kolkata), succeeded in producing and observing electromagnetic waves of much shorter wavelength (25 mm to 5 mm). His experiment , like that of Hertz's was confined to the laboratory. At around the same time, Guglielmo Marconi in Italy followed Hertz's work and succeeded in transmitting electromagnetic waves over distances of many kilometres. Marconi's experiment marks the beginning of the field of communication using electromagnetic waves.

11.3.2 Nature of electromagnetic waves

It can be shown from Maxwell's equations that electric and magnetic fields in an electromagnetic wave are perpendicular to each other, and to the direction of propagation. It appears reasonable , say from our discussion of the displacement current. Consider Fig [11.2]

The electric field inside the plates of the capacitor is directed perpendicular to the plates. The magnetic field this rise to via the displacement current is along the perimeter of a circle parallel to the capacitor plates. So **B** and **E** are perpendicular in this case. This is a general feature.It Fig [11.3] , we show a typical example of a plane electromagnetic wave propagating along the z direction (the fields are shown as a function of the z coordinate, at a given time t).

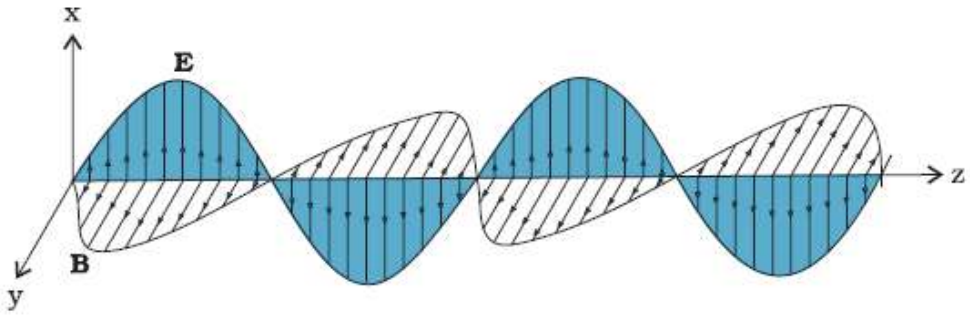


Fig. 11.3 A linearly polarised electromagnetic wave, propagating in the z-direction with the oscillating electric field E along the x-direction and the oscillating magnetic field B along the y-direction

The electric field E_x is along the x-axis, and varies sinusoidally with z, at a given time. The magnetic field B_y is along the y-axis, and again varies sinusoidally with z. The electric and magnetic fields E_x and B_y are perpendicular to each other , and to the direction z of propagation . We can write E_x and B_y as follows :

$$E_x = E_0 \sin(kz - \omega t) \qquad [11.7(a)]$$

$$B_x = B_0 \sin(kz - \omega t) \qquad [11.7(b)]$$

Here k is related to the wave length λ of the wave by the usual equation

$$k = \frac{2\pi}{\lambda} \qquad (11.8)$$

And ω is the angular frequency. K is the magnitude of the wave vector (or propagation vector) **k** and its direction describes the direction of propagation of the wave. The speed of propagation of the wave is (ω/k)

Using Eqs [11.7(a) and (b)] for E_x and B_y and Maxwell's equations , one finds that

$$\omega = ck, where\ , c = 1/\sqrt{\mu_0\varepsilon_0} \qquad [11.9(a)]$$

The relation ω = ck is the standard one for waves . This relation is often written in terms of frequency,

v (=$\omega/2\pi$) and wavelength , $\lambda(= 2\pi/k)$ as

$$2\pi v = c\left(\frac{2\pi}{\lambda}\right) \text{ or } \qquad v\lambda = c \qquad [11.9(b)]$$

It is also seen from Maxwell's equations that the magnitude of the electric and the magnetic fields in an electromagnetic wave are related as

$$B_0 = (E_0/c) \qquad [11.10]$$

We here make remarks on some features of electromagnetic waves. They are self-sustaining oscillations of electric and magnetic fields in free space, or vacuum. They differ from all the other waves we have studied so far, in respect that *no material medium* is involved in the vibrations of the electric and magnetic fields. Sound waves in air are longitudinal waves of compression and rarefaction. Transverse elastic (sound) waves can also propagate in a solid. Which is rigid and that resists shear. Scientists in the nineteenth century were so much used to this mechanical picture that they thought that there must be some medium pervading all space and all matter, which responds to electric and magnetic fields just as any elastic medium does. They called this medium *ether*. They were so convinced of the reality of this medium, that there is even a novel called *The Poison Belt by Sir Arthur Conan Doyle* (the creator of the famous detective *Sherlock Holmes*) where the solar system is supposed to pass through a poisonous region of ether ! We now accept that no such physical medium is needed. The famous experiment of Michelson and Morley in 1887 demolished conclusively the hypothesis of ether. Electric and magnetic fields, oscillating in space and time, can sustain each other in vacuum.

But what if a material medium is actually there? We know that light , we electromagnetic wave, does propagate through glass , for examples. We have seen earlier that the total electric and magnetic fields inside a medium are described in terms of a permittivity ε and a magnetic permeability μ (these describe the factors by which the total fields differ from the external fields). These replace $\varepsilon_0\ and\ \mu_0$ in the description to electric and magnetic fields in Maxwell's equations with the result that in a material medium of permittivity ε and magnetic permeability , the velocity of light becomes,

$$v = \frac{1}{\sqrt{\mu\varepsilon}} \qquad (11.11)$$

Thus, the velocity of light depends on electric and magnetic properties of the medium . We shall see in the next chapter that the refractive index of one medium with respect to the other is equal to the ratio of velocities of light in the two media.

The velocity of electromagnetic waves in free space or vacuum is an important fundamental constant. It has been shown by experiments on electromagnetic waves of different wavelengths that this velocity is the same (independent of wavelength) to within a

few metres per second, out of a value of 3 x 10^8 m/s . The constancy of the velocity of em waves in vacuum is so strongly supported by experiments and the actual value is so well known now that this is used to define a standard of length. Namely, the metre is now defined as the distance travelled by light in vacuum in a time (1/c) seconds = $(2.99792458 \times 10^8)^{-1}$ seconds. This has come about for the following reason. The basic unit of time can be defined very accurately in terms of some atomic frequency, i.e., frequency of light emitted by an atom in a particular process. The basic unit of length is harder to define as accurately in a direct way. Earlier measurement of c using earlier units of length (metre rods, etc). Converged to a value of about 2.9979246 x 10^8 m/s. Since c is such a strongly fixed number, unit of length can be defined in terms of c and the unit of time !

Hertz not only showed the existence of electromagnetic waves, but also demonstrated that the waves, which had wavelength ten million times that of the light waves, could be diffracted, refracted and polarised. Thus, he conclusively established the wave nature of the radiation. Further, he produced stationary electromagnetic waves and determined their wavelength by measuring the distance between two successive nodes. Since the frequency of the wave was known (being equal to the frequency of the oscillator), he obtained the speed of the wave using the formula $v = \nu\lambda$ and found that the waves travelled with the same speed as the speed of light .

The fact that electromagnetic waves are polarised can be easily seen in the response of a portable AM radio to a broadcasting station. If an AM radio has a telescopic antenna., it responds to the electric part of the signal. When the antenna is turned horizontal, the signal will be greatly diminished. Some portable radios have horizontal antenna (usually inside the case of radio), which are sensitive to the magnetic component of the electromagnetic wave. Such a radio must remain horizontal in order to receive the signal. In such cases, response also depends on the orientation of the radio with respect to the station.

Do electromagnetic waves carry energy and momentum like other waves ? Yes, they do. We have seen in chapter 5 that in a region of free space with electric field E, there is an energy density ($\varepsilon_0 E^2/2$). Similarly, as seen in Chapter 9 associated with a magnetic field B is a magnetic energy density ($B^2/2\mu_0$). As electromagnetic wave contains both electric and magnetic fields, there is a non-zero energy density associated with it. Now consider a plane perpendicular to the direction of propagation of the electromagnetic wave (Fig 11.4) . If there are, on this plane, electric charges, they will be set and sustained in motion by the electric and magnetic fields of the electromagnetic waves. This just illustrates the fact that an electromagnetic wave (like other waves) carries energy and momentum. Since it carries momentum, an electromagnetic wave also exerts pressure, called *radiation pressure*.

If the total energy transferred to a surface in time t is U, it can be shown that the magnitude of the momentum delivered to this surface (*for complete absorption*) is,

$$p = \frac{U}{c} \qquad (11.12)$$

When the sun shines on your hand, you feel the energy being absorbed from the electromagnetic waves (your hands get warm). Electromagnetic waves also transfer momentum to your hand but because c is very large, the amount of momentum transferred is extremely small and you do not feel the pressure. In 1903, the American scientists Nicols and Hull succeeded in measuring radiation pressure of visible light and verified Eq (11.12) . It

was found to be of the order of 7 x 10^{-6} N/m^2 . Thus, on a surface of area 10 cm^2 , the force due to radiation is only about 7 x 10^{-9} N.

The great technological importance of electromagnetic waves stems from their capability to carry energy from one le to another .The radio and TV signals from broadcasting stations carry energy. Light carries energy from the sun to the earth , thus making life possible on the earth.

11.4 ELECTROMAGNETIC SPECTRUM

At the time Maxwell predicted the existence of electromagnetic waves , the only familiar electromagnetic waves were the visible light waves. The existence of ultraviolet and infrared waves was barely established. By the end of the nineteenth century, X-rays and gamma rays had also been discovered. We now know that, electromagnetic waves include visible light waves, X-rays, gamma rays, radio waves, and microwaves, ultraviolet and infrared waves. The classification of em waves according to frequency is the electromagnetic spectrum (Fig 11.5). *There is no sharp division between one kind of wave and the next.* The classification is based roughly on how the waves are produced and / or detected.

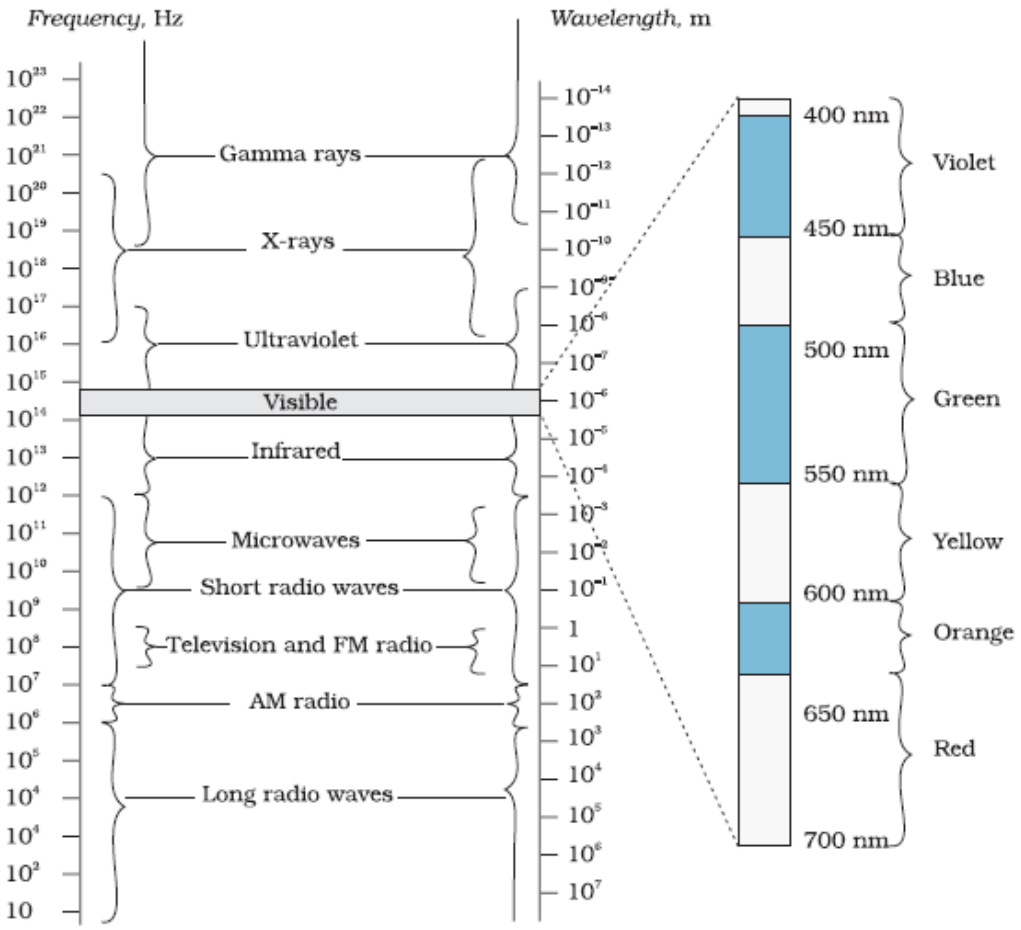


Fig. 11.4 The electromagnetic spectrum, with common names for various part of it. The various regions do not have sharply defined boundaries

We briefly describe these different types of electromagnetic waves, in order of decreasing wavelengths.

11.4.1 Radio waves

Radio waves are produced by the accelerated motion charges in conducting wires. They are used in ratio and television communication systems. They are generally in the frequency range from 500 kHz to about 1000 MHz. The AM(amplitude modulated) band is from 530 kHz to 1710 kHz. Higher frequencies upto 54 MHz are used for short wave bands. TV waves range from 54 MHz to 890 MHz. The FM(frequency modulated) radio band extends from 88 MHz to 108 MHz. Cellular phones use radio waves to transmit voice communication in the ultrahigh frequency (UHF) band. How these waves are transmitted and received is described in Chapter**.**

11.4.2 Microwaves

Microwaves (short-wavelength radio waves), with frequencies in the gigahertz (GHz) range, and are produced by special vacuum tubes (called klystrons, magnetrons and Gunn diodes). Due to their short wavelengths, they are suitable for the radar systems used in aircraft navigation. Radar also provides the basis for the speed guns used to time fast balls, tennis-serves, and automobiles. Microwave ovens are an interesting domestic application of these waves. In such ovens, the frequency of the microwaves is selected to match the resonant frequency of water molecules so that energy from the waves is transferred efficiently to the kinetic energy of the molecules. This raises the temperature of any food containing water.

MICROWAVE OVEN

The spectrum of *electromagnetic radiation* contains a part known as *microwaves.* These waves have frequency and energy smaller than visible light and wavelength larger than it. What is the principle of a microwave oven and how does it work?

Our objective is to cook food or warm it up. All food items such as fruit, vegetables, meat, cereals, etc., contain water as a constituent. Now, what does it mean when we say that a certain object has become warmer? When the temperature of a body rises, the energy of the random motion of atoms and molecules increases and the molecules travel or vibrate or rotate with higher energies. The frequency of rotation of water molecules is about 300 crore hertz, which is 3 gigahertz (GHz). If water receives microwaves of this frequency, its molecules absorb this radiation, which is equivalent to heating up water. These molecules share this energy with neighbouring food molecules, heating up the food.

One should use porcelain vessels and not metal containers in a microwave oven because of the danger of getting a shock from accumulated electric charges. Metals may also melt from heating. The porcelain container remains unaffected and cool, because its large molecules vibrate and rotate with much smaller frequencies, and thus cannot absorb microwaves. Hence, they do not get heated up.

Thus, the basic principle of a microwave oven is to generate microwave radiation of appropriate frequency in the working space of the oven where we keep food. This way energy is not wasted in heating up the vessel. In the conventional heating method, the vessel on the burner gets heated first, and then the food inside gets heated because of transfer of energy from the vessel. In the microwave oven, on the other hand, energy is directly delivered to water molecules which is shared by the entire food.

11.4.3 Infrared waves

Infrared waves are produced by hot bodies and molecules. This band lies adjacent to the low-frequency or long-wave length end of the visible spectrum. Infrared waves are sometimes referred to as *heat waves*. This is because water molecules present in most materials readily absorb infrared waves (many other molecules, for example, CO_2, NH_3, also absorb infrared waves). After absorption, their thermal motion increases, that is, they heat up and heat their surroundings. Infrared lamps are used in physical therapy. Infrared radiation also plays an important role in maintaining the earth's warmth or average temperature through the greenhouse effect. Incoming visible light (which passes relatively easily through the atmosphere) is absorbed by the earth's surface and reradiated as infrared (longer wavelength) radiations. This radiation is trapped by greenhouse gases such as carbon dioxide and water vapour. Infrared detectors are used in Earth satellites, both for military purposes and to observe growth of crops. Electronic devices (for example semiconductor light emitting diodes) also emit infrared and are widely used in the remote switches of household electronic systems such as TV sets, video recorders and hi-fi-systems.

11.4.4 Visible rays

It is the most familiar form of electromagnetic waves. It is the part of the spectrum that is detected by the human eye. It runs from about 4×10^{14} Hz to about 7×10^{14} Hz or a wavelength range of about 700-400 nm. Visible light emitted or reflected from objects around us provides us information about the world. Our eyes are sensitive to this range of wavelengths. Different animals are sensitive to different range of wavelengths. For example, snakes can detect infrared waves, and the 'visivle' range of many insects extends well into the ultraviolet.

11.4.5 Ultraviolet rays

It covers wavelengths ranging from about 4×10^{-7} m (400 nm) down to 6×10^{-10}m (0.6 nm) . Ultraviolet (UV) radiation is produced by special lamps and very hot bodies The sun is an important source of ultraviolet light. But fortunately , most of it is absorbed in the ozone layer in the atmosphere at an altitude of about 40-50 km. UV light in large quantities has harmful effects on humans. Exposure to UV radiation induces the production of more melanin, causing tanning of the skin. UV radiation is absorbed by ordinary glass. Hence, one cannot get tanned or sunburn through glass windows.

Welders wear special glass goggles or face masks with windows to protect their eyes from large amount of UV produced By welding arcs. Due to its shorter wavelengths, UV radiaions can be focussed into very narrow beams for hoigh precision applications such as LASIK (*Laser assisted in situ keratomileusis*) eye surgery. UV lamps are used to kill germs in water puriiers. Ozone layer in the atmosphere plays a protective role, and hence its depletion by chlorofluorocarbons (CFCs) gas (such as freon) is a matter of international concern.

11.4.6 X-rays

Beyond the UV region of the electromagnetic spectrum lies the X-ray regions. We are familiar with X-rays because of its medical applications. It covers wavelengths from about 10^{-8} m (10 nm) down to 10^{-13} m (10^{-4} nm) . One common way to generate X-rays is to bombard a metal target by high energy electrons. X-rays are used as a diagnostic tool in medicine and as a treatment for certain forms of cancer. Because X-rays damage or destroy living tissue and organisms, care must be taken to avoid unnecessary or over exposure.

11.4.7 Gamma rays

They lie in the upper frequency range of the electromagnetic spectrum and have wavelengths of from about 10^{-10} m to less than 10^{-14} m. This high frequency radiation is produced in nuclear reactions and also emitted by radioactive nuclei. They are used in medicine to destroy cancer cells.

Table 11.1 summarises different types of electromagnetic waves, their production and detections. As mentioned earlier, the demarcation between different regions is not sharp and there are overlaps.

Table 11.1 Different types of electromagnetic waves

Type	Wavelength range	Production	Detection
Radio	> 0.1 m	Rapid acceleration and decelerations of electrons in aerials	Receiver's aerials
Microwave	0.1m to 1 mm	Klystron valve or magnetron valve	Point contact diodes
Infra-red	1mm to 700 nm	Vibration of atoms and molecules	Thermopiles Bolometer, Infrared photographic film
Light	700 nm to 400 nm	Electrons in atoms emit light when they move from one energy level to a lower energy level	The eye Photocells Photographic film
Ultraviolet	400 nm to 1nm	Inner shell electrons in atoms moving from one energy level to a lower level	Photocells Photographic film
X-rays	1nm to 10^{-3} nm	X-ray tubes or inner shell electrons	Photographic film Geiger tubes Ionisation chamber
	Gamma rays	$<10^{-3}$ nm	Radioactive decay of the nucleus-do

SUMMARY

1. Maxwell found an inconsistency in the Ampere's law and suggested the existence of an additional current, called displacement current, to remove this inconsistency. This displacement current is due to time-varying electric field and is given by

$$i_d = \varepsilon_0 \frac{d\phi_E}{dt}$$

and acts as a source of magnetic field in exactly the same way as conduction current.

2. An accelerating charge produces electromagnetic waves. An electric charge oscillating harmonically with frequency ν, produces electromagnetic waves of the same frequency ν. An electric dipole is a basic source of electromagnetic waves.
3. Electromagnetic waves with wavelength of the order of a few metres were first produced and detected in the laboratory by Hertz in 1887. He thus verified a basic prediction of Maxwell's equations.
4. Electric and magnetic fields oscillate sinusoidally in space and time in an electromagnetic wave. The oscillating electric and magnetic fields, **E** and **B** are perpendicular to each other, and to the direction of propagation of the electromagnetic wave. For a wave of frequency ν, wavelength λ, propagating along *z*-direction, we have

$$\begin{aligned} E &= E_x(t) = E_0 \sin(kz - \omega t) \\ &= E_0 \sin\left[2\pi\left(\frac{z}{\lambda} - \nu t\right)\right] \\ &= E_0 \sin\left[2\pi\left(\frac{z}{\lambda} - \frac{t}{T}\right)\right] \\ B &= B_y(t) = B_0 \sin(kz - \omega t) \\ &= B_0 \sin\left[2\pi\left(\frac{z}{\lambda} - \nu t\right)\right] \\ &= B_0 \sin\left[2\pi\left(\frac{z}{\lambda} - \frac{t}{T}\right)\right] \end{aligned}$$

They are related by $E_0 / B_0 = c$.

5. The speed *c* of electromagnetic wave in vacuum is related to μ0 and ε0 (the free space permeability and permittivity constants) as follows:
$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}}$. The value of *c* equals the speed of light obtained from optical measurements. Light is an electromagnetic wave; c is, therefore, also the speed of light. Electromagnetic waves other than light also have the same velocity c in free space. The speed of light or of electromagnetic waves in a material medium is given by $v = 1/\sqrt{\mu \varepsilon}$ where μ is the permeability of the medium and ε its permittivity.
6. Electromagnetic waves carry energy as they travel through space and this energy is shared equally by the electric and magnetic fields. Electromagnetic waves transport momentum as well. When these waves strike a surface, a pressure is exerted on the surface. If total energy transferred to a surface in time *t* is *U*, total momentum delivered to this surface is $p = U/c$.
7. The spectrum of electromagnetic waves stretches, in principle, over an infinite range of wavelengths. Different regions are known by different names; γ-rays, X-rays, ultraviolet rays, visible rays, infrared rays, microwaves and radio waves in order of increasing wavelength from 10^{-2} Å or 10^{-12} m to 10^{6}m.

They interact with matter via their electric and magnetic fields which set in oscillation charges present in all matter. The detailed interaction and so the mechanism of absorption, scattering, etc., depend on the wavelength of the electromagnetic wave, and the nature of the atoms and molecules in the medium.

VERY SHORT ANSWER QUESTIONS (2 marks)

1. What is the average wavelength of X-rays?
2. Give any two uses of infrared rays.
3. If the wavelength of electromagnetic radiation is doubled, what happens to the energy of photon?
4. What is the principle of production of electromagnetic waves?
5. What is the ratio of speed of infrared rays and ultraviolet rays in vacuum?
6. What is the relation between the amplitudes of the electric and magnetic field in free space for an electromagnetic wave?
7. What are the applications of microwaves?
8. Microwaves are used in Radars, why?

SHORT ANSWER QUESTIONS (4 Marks)

1. What does an electromagnetic wave consists of ? On what factors does its velocity in vacuum depend?
2. What is Greenhouse effect and its contribution towards the surface temperature of earth?

LONG ANSWER QUESTIONS (8 Marks)

1. State six characteristics of electromagnetic waves. What is Greenhouse effect?

CHAPTER 12

DUAL NATURE OF RADIATION AND MATTER

12.1 Introduction

The Maxwell's equations of electromagnetism and Hertz experiments on the generation and detection of electromagnetic waves in 1887 strongly established the wave nature of light. Towards the same period at the end of 19th century, experimental investigations on conduction of electricity (electric discharge) through gases at low pressure in a discharge tube led to many historic discoveries. The discovery of X-rays was by Roentgen in 1895, and of electron by J.J. Thomson in 1897, were important milestones in the understanding of atomic structure. It was found that at sufficiently low pressure of about 0.001 mm of mercury column, a discharge took place between the two electrodes on applying the electric field to the gas in the discharge tube. A fluorescent glow appeared on the glass opposite to cathode. The colour of glow of the glass depended on the type of glass, it being yellowish-green for soda glass. The cause of this fluorescence was attributed to the radiation which appeared to be coming from the cathode. These *cathode rays* were discovered, in 1870, by William Crookes who later, in 1879, suggested that these rays consisted of streams of fast moving negatively charged particles. The British physicist J. J. Thomson (1856-1940) confirmed this hypothesis. By applying mutually perpendicular electric and magnetic fields across the discharge tube, J. J. Thomson was the first to determine experimentally the speed and the specific charge [charge to mass ratio (e/m)] of the cathode ray particles. They were found to travel with speeds ranging from about 0.1 to 0.2 times the speed of light ($3 \times l0^8$ m/s). The presently accepted value of *e/m* is 1.76×10^{11} C/kg. Further, the value of *e/m* was found to be independent of the nature of the material/metal used as the cathode (emitter), or the gas introduced in the discharge tube. This observation suggested the universality of the cathode ray particles.

Around the same time, in 1887, it was found that certain metals, when irradiated by ultraviolet light, emitted negatively charged particles having small speeds. Also, certain metals when heated to a high temperature were found to emit negatively charged particles. The value of *e/ m* of these particles were found to be the same as that for cathode ray particles. These observations thus established that all these particles, although produced under different conditions, were identical in nature. J.J. Thomson, in 1897, named these particles as *electrons,* and suggested that they were fundamental, universal constituents of matter. For his epoch-making discovery of electron, through his theoretical and experimental investigations on conduction of electricity by gasses, he was awarded the Nobel Prize in Physics in 1906. In 1913, the American physicist R.A. Millikan (1868-1953) performed the pioneering oil-drop experiment for the precise measurement of the charge on an electron. He found that the charge on an oil-droplet was always an integral multiple of an elementary charge, 1.602 x 10"[19] C. Millikan's experiment established that *electric charge is quantised.* From the values of charge (e) and specific charge (e/m), the mass (m) of the electron could be determined.

12.2 Electron Emission

We know that metals have free electrons (negatively charged particles) that are responsible for their conductivity. However, the free electrons cannot normally escape out of the metal surface. If an electron attempts to come out of the metal, the metal surface acquires a positive charge and pulls the electron back to the metal. The free electron is thus held inside the metal surface by the attractive forces of the ions. Consequently, the electron can come out of the metal surface only if it has got sufficient energy to overcome the attractive pull. A certain minimum amount of energy is required to be given to an electron to pull it out from

the surface of the metal. This minimum energy required by an electron to escape from the metal surface is called the *work function* of the metal. It is generally denoted by ϕ_0 and measured in eV (electron volt). One electron volt is the energy gained by an electron when it has been accelerated by a potential difference of 1 volt, so that 1 eV = 1.602×10^{19} J.

This unit of energy is commonly used in atomic and nuclear physics. The work function (ϕ_O) depends on the properties of the metal and the nature of its surface. The values of work function of some metals are given in Table 12.1. These values are approximate as they are very sensitive to surface impurities.

Note from Table 12.1 that the work function of platinum is the highest (ϕ_0 = 5.65 eV) while it is the lowest (ϕ_0 = 2.14 eV) for caesium.

The minimum energy required for the electron emission from the metal surface can be supplied to the free electrons by any one of the following physical processes:

Table 12.1 Work Functions of some metals

Metal	Work function ϕ_0 (eV)	Metal	Work function ϕ_0 (eV)
Cs	2.14	A1	4.28
K	2.30	Hg	4.49
Na	2.75	Cu	4.65
Ca	3.20	Ag	4.70
Mo	4.17	Ni	5.15
Pb	4.25	Pt	5.65

(i) *Thermionic emission:* By suitably heating, sufficient thermal energy can be imparted to the free electrons to enable them to come out of the metal.

(ii) *Field emission*: By applying a very strong electric field (of the order of 10^8 V m"[1]) to a metal, electrons can be pulled out of the metal, as in a spark plug.

(iii) *Photo-electric emission*: When light of suitable frequency illuminates a metal surface, electrons are emitted from the metal surface. These photo (light)-generated electrons are called *photoelectrons.*

12.3 Photoelectric Effect

12.3.1 Hertz's observations

The phenomenon of photoelectric emission was discovered in 1887 by Heinrich Hertz (1857-1894), during his electromagnetic wave experiments. In his experimental investigation on the production of electromagnetic waves by means of a spark discharge, Hertz observed that high voltage sparks across the detector loop were enhanced when the emitter plate was illuminated by ultraviolet light from an arc lamp.

Light shining on the metal surface somehow facilitated the escape of free, charged particles which we now know as electrons. When light falls on a metal surface some electrons near the surface absorb enough energy from the incident radiation to overcome the attraction of the positive ions in the material of the surface. After gaining sufficient energy from the incident light, the electrons escape from the surface of the metal into the surrounding space.

12.3.2 Hallwachs' and Lenard's observations

Wilhelm Hallwachs and Philipp Lenard investigated the phenomenon of photoelectric emission in detail during 1886-1902.

Lenard (1862-1947) observed that when ultraviolet radiations were allowed to fall on the emitter plate of an evacuated glass tube enclosing two electrodes (metal plates), current flows in the circuit (Fig. 12.1). As soon as the ultraviolet radiations were stopped, the current

flow also stopped. These observations indicate that when ultraviolet radiations fall on the emitter plate C, electrons are ejected from it which is attracted towards the positive, collector plate A by the electric field. The electrons flow through the evacuated glass tube, resulting in the current flow. Thus, light falling on the surface of the emitter causes current in the external circuit. Hallwachs and Lenard studied how this photo current varied with collector plate potential, and with frequency and intensity of incident light.

Hallwachs, in 1888, undertook the study further and connected a negatively charged zinc plate to an electroscope. He observed that the zinc plate lost its charge when it was illuminated by ultraviolet light. Further, the uncharged zinc plate became positively charged when it was irradiated by ultraviolet light. Positive charge on a positively charged zinc plate was found to be further enhanced when it was illuminated by ultraviolet light. From these observations he concluded that negatively charged particles were emitted from the zinc plate under the action of ultraviolet light.

After the discovery of the electron in 1897, it became evident that the incident light causes electrons to be emitted from the emitter plate. Due to negative charge, the emitted electrons are pushed towards the collector plate by the electric field. Hallwachs and Lenard also observed that when ultraviolet light fell on the emitter plate, no electrons were emitted at all when the frequency of the incident light was smaller than a certain minimum value, called the *threshold frequency*. This minimum frequency depends on the nature of the material of the emitter plate.

It was found that certain metals like zinc, cadmium, magnesium, etc., responded only to ultraviolet light, having short wavelength, to cause electron emission from the surface. However, some alkali metals such as lithium, sodium, potassium, caesium and rubidium were sensitive even to visible light. All these *photosensitive substances* emit electrons when they are illuminated by light. After the discovery of electrons, these electrons were termed as *photoelectrons.* The phenomenon is called *photoelectric effect*.

12.4 Experimental Study of Photoelectric Effect

Figure 12.1 depicts a schematic view of the arrangement used for the experimental study of the photoelectric effect. It consists of an evacuated glass/quartz tube having a photosensitive plate C and another metal plate A. Monochromatic light from the source S of sufficiently short wavelength passes through the window W and falls on the photosensitive plate C (emitter). A transparent quartz window is sealed on to the glass tube, which permits ultraviolet radiation to pass through it and irradiate the photosensitive plate C. The electrons are emitted by the plate C and are collected by the plate A (collector), by the electric field created by the battery. The battery maintains the potential difference between the plates C and A, that can be varied. The polarity of the plates C and A can be reversed by a commutator. Thus, the plate A can be maintained at a desired positive or negative potential with respect to emitter C. When the collector plate A is positive with respect to the emitter plate C, the electrons are attracted to it. The emission of electrons causes flow of electric current in the circuit. The potential difference between the emitter and collector plates is measured by a voltmeter (V) whereas the resulting photo current flowing in the circuit is measured by a micro-ammeter (nA). The photoelectric current can be increased or decreased by varying the potential of collector plate A with respect to the emitter plate C. The intensity and frequency of the incident light can be varied, as can the potential difference V between the emitter C and the collector A.

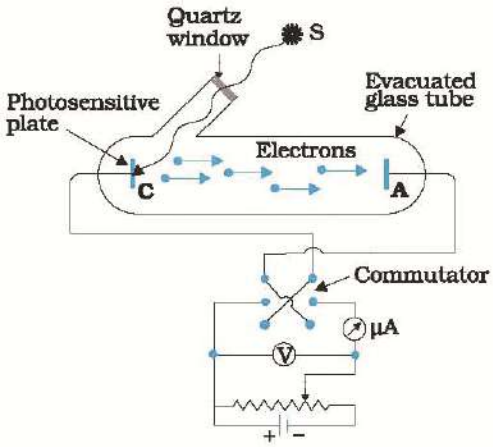


Fig. 12.1 Experimental arrangement for study of photoelectric effect.

We can use the experimental arrangement of Fig. 12.1 to study the variation of photocurrent with (a) intensity of radiation, (b) frequency of incident radiation, (c) the potential difference between the plates A and C, and (d) the nature of the material of plate C. Light of different frequencies can be used by putting appropriate coloured filter or coloured glass in the path of light falling on the emitter C. The intensity of light is varied by changing the distance of the light source from the emitter.

12.4.1 Effect of intensity of light on photocurrent

The collector A is maintained at a positive potential with respect to emitter C so that electrons ejected from C are attracted towards collector A. Keeping the frequency of the incident radiation and the accelerating potential fixed, the intensity of light is varied and the resulting photoelectric current is measured each time. It is found that the photocurrent increases linearly with intensity of incident light as shown graphically in Fig. 12.2. The photocurrent is directly proportional to the number of photoelectrons emitted per second. This implies that *the number of photoelectrons emitted per second is directly proportional to the intensity of incident radiation.*

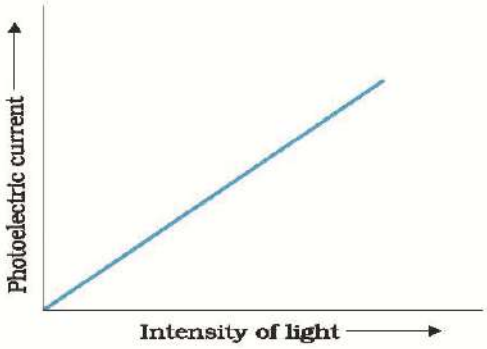


Fig. 12.2 Variation of Photoelectric current with intensity of light.

12.4.2 Effect of potential on photoelectric current

We first keep the plate A at some positive accelerating potential with respect to the plate C and illuminate the plate C with light of fixed frequency v and fixed intensity I,. We next vary the positive potential of plate A gradually and measure the resulting photocurrent each time. It is found that the photoelectric current increases with increase in accelerating (positive) potential. At some stage, for a certain positive potential of plate A, all the emitted electrons are collected by the plate A and the photoelectric current becomes maximum or saturates. If we increase the accelerating potential of plate A further, the photocurrent does not increase. This maximum value of the photoelectric current is called *saturation current*

Saturation current corresponds to the case when all the photoelectrons emitted by the emitter plate C reach the collector plate A.

We now apply a negative (retarding) potential to the plate A with respect to the plate C and make it increasingly negative gradually. When the polarity is reversed, the electrons are repelled and only the most energetic electrons are able to reach the collector A. The photocurrent is found to decrease rapidly until it drops to zero at a certain sharply defined, critical value of the negative potential V^, on the plate A. For a particular frequency of incident radiation, *the minimum negative (retarding) potential V_0 given to the plate A for which the photocurrent stops or becomes zero is called the cut-off or stopping potential.*

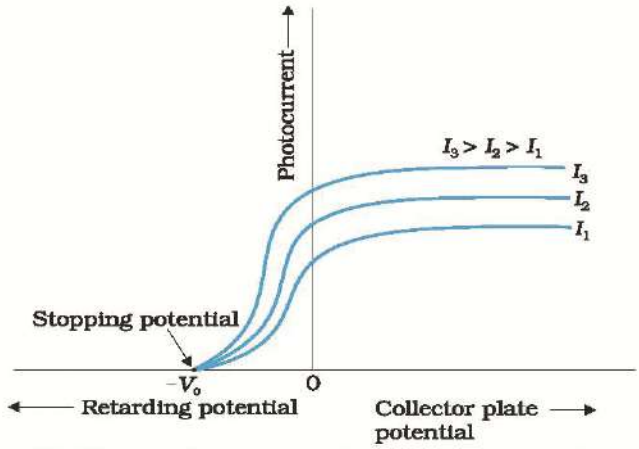


Fig. 12.3 Variation of photocurrent with collector plate potential for different intensity of incident radiation.

The interpretation of the observation in terms of photoelectrons is straightforward. All the photoelectrons emitted from the metal do not have the same energy. Photoelectric current is zero when the stopping potential is sufficient to repel even the most energetic photoelectrons, with the maximum kinetic energy (K_{max}), so that

$$K_{max} = eV_0 \qquad (12.1)$$

We can now repeat this experiment with incident radiation of the same frequency but of higher intensity I_2 and I_3 ($I_3 > I_2 > I$,). We note that the saturation currents are now found to be at higher values. This shows that more electrons are being emitted per second, proportional to the intensity of incident radiation. But the stopping potential remains the same as that for the incident radiation of intensity I,, as shown graphically in Fig. 12.3. Thus, *for a given frequency of the incident radiation, the stopping potential is independent of its intensity.* In other words, the maximum kinetic energy of photoelectrons depends on the light source and the emitter plate material, but is independent of intensity of incident radiation.

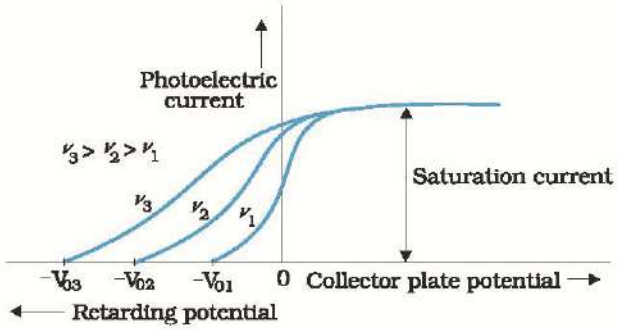


Fig. 12.4 Variation of photoelectric current with collector plate potential for different frequencies of incident radiation.

12.4.3 Effect of frequency of incident radiation on stopping potential

We now study the relation between the frequency v of the incident radiation and the stopping potential V_0. We suitably adjust the same intensity of light radiation at various frequencies and study the variation of photocurrent with collector plate potential. The resulting variation is shown in Fig.12.4. We obtain different values of stopping potential but the same value of the saturation current for incident radiation of different frequencies. The energy of the emitted electrons depends on the frequency of the incident radiations. The stopping potential is more negative for higher frequencies of incident radiation. Note from Fig. 12.4 that the stopping potentials are in the order $V_{03} > V_{02} > V_{01}$ if the frequencies are in the order $v_3 > v_2 > v$,. This implies that greater the frequency of incident light, greater is the maximum kinetic energy of the photoelectrons. Consequently, we need greater retarding potential to stop them completely. If we plot a graph between the frequency of incident radiation and the corresponding stopping potential for different metals we get a straight line, as shown in Fig.12.5.

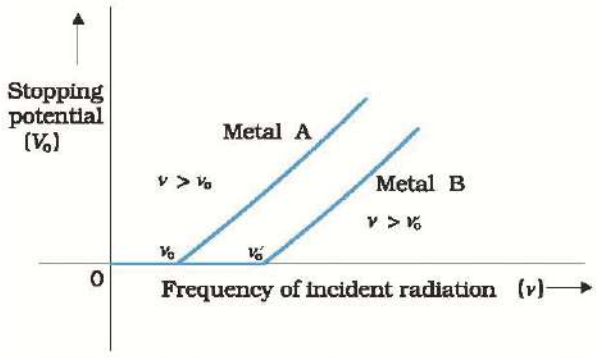


Fig. 12.5 Variation of stopping potential V_0 with frequency v of incident radiation for a given photosensitive material.

The graph shows that

(i) the stopping potential V_0 varies linearly with the frequency of incident radiation for a given photosensitive material.

(ii) there exists a certain minimum cut-off frequency v_0 for which the stopping potential is zero.

These observations have two implications:

(i) The maximum kinetic energy of the photoelectrons varies linearly with the frequency of incident radiation, but is independent of its intensity.

(ii) For a frequency v of incident radiation, lower than the cut-off frequency v_0, no photoelectric emission is possible even if the intensity is large.

This minimum, cut-off frequency v_0, is called the *threshold frequency.* It is different for different metals.

Different photosensitive materials respond differently to light. Selenium is more sensitive than zinc or copper. The same photosensitive substance gives different response to light of different wavelengths. For example, ultraviolet light gives rise to photoelectric effect in copper while green or red light does not.

Note that in all the above experiments, it is found that, if frequency of the incident radiation exceeds the threshold frequency, the photoelectric emission starts instantaneously without any apparent time lag, even if the incident radiation is very dim. It is now known that emission starts in a time of the order of 10^{-9} s or less.

We now summarise the experimental features and observations described in this section.

(i) For a given photosensitive material and frequency of incident radiation (above the threshold frequency), the photoelectric current is directly proportional to the intensity of incident light (Fig. 12.2).

(ii) For a given photosensitive material and frequency of incident radiation, saturation current is found to be proportional to the intensity of incident radiation whereas the stopping potential is independent of its intensity (Fig. 12.3).

(iii) For a given photosensitive material, there exists a certain minimum cut-off frequency of the incident radiation, called the *threshold frequency,* below which no emission of photoelectrons takes place, no matter how intense the incident light is. Above the threshold frequency, the stopping potential or equivalently the maximum kinetic energy of the emitted photoelectrons increases linearly with the frequency of the incident radiation, but is independent of its intensity (Fig. 12.5).

(iv) The photoelectric emission is an instantaneous process without any apparent time lag ($\sim 10^{-9}$ s or less), even when the incident radiation is made exceedingly dim.

12.5 Photoelectric Effect and Wave Theory of Light

The wave nature of light was well established by the end of the nineteenth century. The phenomena of interference, diffraction and polarisation were explained in a natural and satisfactory way by the wave picture of light. According to this picture, light is an electromagnetic wave consisting of electric and magnetic fields with continuous distribution of energy over the region of space over which the wave is extended. Let us now see if this wave picture of light can explain the observations on photoelectric emission given in the previous section.

According to the wave picture of light, the free electrons at the surface of the metal (over which the beam of radiation falls) absorb the radiant energy continuously. The greater the intensity of radiation, the greater is the amplitude of electric and magnetic fields. Consequently, the greater the intensity, the greater should be the energy absorbed by each electron. In this picture, the maximum kinetic energy of the photoelectrons on the surface is then expected to increase with increase in intensity. Also, no matter what the frequency of radiation is, a sufficiently intense beam of radiation (over sufficient time) should be able to impart enough energy to the electrons, so that they exceed the minimum energy needed to escape from the metal surface. A threshold frequency, therefore, should not exist. These expectations of the wave theory directly contradict observations (i), (ii) and (iii) given at the end of sub-section 12.4.3.

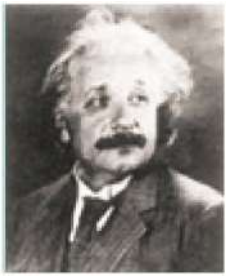
Further, we should note that in the wave picture, the absorption of energy by electron takes place continuously over the entire wave front of the radiation. Since a large number of electrons absorb energy, the energy absorbed per electron per unit time turns out to be small. Explicit calculations estimate that it can take hours or more for a single electron to pick up sufficient energy to overcome the work function and come out of the metal. This conclusion is again in striking contrast to observation (iv) that the photoelectric emission is instantaneous. In short, the wave picture is unable to explain the most basic features of photoelectric emission.

12.6 Einstein's Photoelectric Equation: Energy Quantum of Radiation

In 1905, Albert Einstein (1879-1955) proposed a radically new picture of electromagnetic radiation to explain photoelectric effect. In this picture, photoelectric emission does not take place by continuous absorption of energy from radiation. Radiation energy is built up of discrete units - the so called *quanta of energy of radiation.* Each

quantum of radiant energy has energy *hv*, where *h* Is Planck's constant and v the frequency of light. In photoelectric effect, an electron absorbs a quantum of energy (hv) of radiation. If this quantum of energy absorbed exceeds the minimum energy needed for the electron to escape from the metal surface (work function ϕ_0), the electron Is emitted with maximum kinetic energy

$$K_{max} = h\nu - \phi_0 \qquad (12.2)$$



Albert Einstein (1879 - 1955) Einstein, one of the greatest physicists of all time, was born in Ulm, Germany. In 1905, he published three path- breaking papers. In the first paper, he introduced the notion of light quanta (now called photons) and used it to explain the features of photoelectric effect. In the second paper, he developed a theory of Brownian motion, confirmed experimentally a few years later and provided a convincing evidence of the atomic picture of matter. The third paper gave birth to the special theory of relativity. In 1916, he published the general theory of relativity. Some of Einstein's most significant later contributions are: the notion of stimulated emission introduced in an alternative derivation of Planck's blackbody radiation law, static model of the universe which started modern cosmology, quantum statistics of a gas of massive bosons, and a critical analysis of the foundations of quantum mechanics. In 1921, he was awarded the Nobel Prize in physics for his contribution to theoretical physics and the photoelectric effect.

More tightly bound electrons will emerge with kinetic energies less than the maximum value. Note that the intensity of light of a given frequency is determined by the number of photons incident per second. Increasing the intensity will increase the number of emitted electrons per second. However, the maximum kinetic energy of the emitted photoelectrons is determined by the energy of each photon.

Equation (12.2) is known as *Einstein's photoelectric equation.* We now see how this equation accounts in a simple and elegant manner all the observations on photoelectric effect given at the end of sub-section 12.4.3.

- According to Eq. (12.2), K_{max} depends linearly on v, and is independent of intensity of radiation, in agreement with observation. This has happened because in Einstein's picture, photoelectric effect arises from the absorption of a single quantum of radiation by a single electron. The intensity of radiation (that is proportional to the number of energy quanta per unit area per unit time) is irrelevant to this basic process.
- Since K_{max} must be non-negative, Eq. (12.2) implies that photoelectric emission is possible only if

  $$h\nu > \phi_0$$

  or $\nu > \nu_0$, where

  $$\nu_0 = \frac{\phi_0}{h} \qquad (12.3)$$

  Equation (12.3) shows that the greater the work function ϕ_0, the higher the minimum or threshold frequency ν_0 needed to emit photoelectrons. Thus, there exists a threshold frequency ν_0 (= ϕ_0/h) for the metal surface, below which no photoelectric emission is possible, no matter how intense the incident radiation may be or how long it falls on the surface.
- In this picture, intensity of radiation as noted above, is proportional to the number of energy quanta per unit area per unit time. The greater the number of energy quanta

available, the greater is the number of electrons absorbing the energy quanta and greater, therefore, is the number of electrons coming out of the metal (for $v > v_0$). This explains why, for $v > v_0$, photoelectric current is proportional to intensity.

- In Einstein's picture, the basic elementary process involved in photoelectric effect is the absorption of a light quantum by an electron. This process is instantaneous. Thus, whatever may be the intensity i.e., the number of quanta of radiation per unit area per unit time, photoelectric emission is instantaneous. Low intensity does not mean delay in emission, since the basic elementary process is the same. Intensity only determines how many electrons are able to participate in the elementary process (absorption of a light quantum by a single electron) and, therefore, the photoelectric current.

Using Eq. (12.1), the photoelectric equation, Eq. (12.2), can be written as

$$e V_0 = h\nu - \phi_0; \text{ for } \nu \geq \nu_0$$

$$\text{or} \quad V_0 = \left(\frac{h}{e}\right)\nu - \frac{\phi_0}{e} \qquad (12.4)$$

This is an important result. It predicts that the V_0 versus v curve is a straight line with slope = (h/e), independent of the nature of the material. During 1906-1916, Millikan performed a series of experiments on photoelectric effect, aimed at disproving Einstein's photoelectric equation. He measured the slope of the straight line obtained for sodium, similar to that shown in Fig. 12.5. Using the known value of e, he determined the value of Planck's constant *h*. This value was close to the value of Planck's constant (=6.626 × 10^{-34} Js) determined in an entirely different context. In this way, in 1916, Millikan proved the validity of Einstein's photoelectric equation, instead of disproving it.

The successful explanation of photoelectric effect using the hypothesis of light quanta and the experimental determination of values of hand ϕ_0, in agreement with values obtained from other experiments, led to the acceptance of Einstein's picture of photoelectric effect. Millikan verified photoelectric equation with great precision, for a number of alkali metals over a wide range of radiation frequencies.

12.7 Particle Nature of Light: The Photon

Photoelectric effect thus gave evidence to the strange fact that light in interaction with matter behaved as if it was made of quanta or packets of energy, each of energy $h\nu$.

Is the light quantum of energy to be associated with a particle? Einstein arrived at the important result, that the light quantum can also be associated with momentum ($h\nu/c$). A definite value of energy as well as momentum is a strong sign that the light quantum can be associated with a particle. This particle was later named *photon*. The particle-like behaviour of light was further confirmed, in 1924, by the experiment of A.H. Compton (1892-1962) on scattering of X-rays from electrons. In 1921, Einstein was awarded the Nobel Prize in Physics for his contribution to theoretical physics and the photoelectric effect. In 1923, Millikan was awarded the Nobel Prize in physics for his work on the elementary charge of electricity and on the photoelectric effect.

We can summarise the photon picture of electromagnetic radiation as follows:

(i) In interaction of radiation with matter, radiation behaves as If It Is made up of particles called photons.

(ii) Each photon has energy E (= $h\nu$) and momentum p (= $h\nu/c$), and speed c, the speed of light.

(iii) All photons of light of a particular frequency v, or wavelength A, have the same energy E (= $h\nu = hc/\lambda$) and momentum p (= $h\nu/c = h/\lambda$), whatever the Intensity of radiation

may be. By increasing the intensity of light of given wavelength, there is only an increase in the number of photons per second crossing a given area, with each photon having the same energy. Thus, photon energy is independent of intensity of radiation.

(iv) Photons are electrically neutral and are not deflected by electric and magnetic fields.

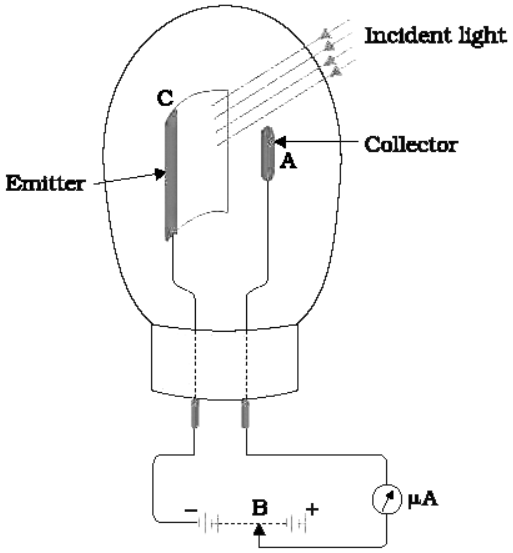
(v) In a photon-particle collision (such as photon-electron collision), the total energy and total momentum are conserved. However, the number of photons may not be conserved in a collision. The photon may be absorbed or a new photon may be created.

PHOTOCELL

A photocell is a technological application of the photoelectric effect. It is a device whose electrical properties are affected by light. It is also sometimes called an electric eye. A photocell consists of a semi-cylindrical photo-sensitive metal plate C (emitter) and a wire loop A (collector) supported in an evacuated glass or quartz bulb. It is connected to the external circuit having a high-tension battery B and microammeter (juA) as shown in the Figure. Sometimes, instead of the plate C, a thin layer of photosensitive material is pasted on the inside of the bulb. A part of the bulb is left clean for the light to enter it.

When light of suitable wavelength falls on the emitter C, photoelectrons are emitted. These photoelectrons are drawn to the collector A. Photocurrent of the order of a few microampere can be normally obtained from a photo cell.

A photocell converts a change in intensity of illumination into a change in photocurrent. This current can be used to operate control systems and in light measuring devices. A photocell of lead sulphide sensitive to infrared radiation is used in electronic ignition circuits.



A photo cell

In scientific work, photo cells are used whenever it is necessary to measure the intensity of light. Light meters in photographic cameras make use of photo cells to measure the intensity of incident light. The photocells, inserted in the door light electric circuit, are used as automatic door opener. A person approaching a doorway may interrupt a light beam which is incident on a photocell. The abrupt change in photocurrent may be used to start a motor which opens the door or rings an alarm. They are used in the control of a counting device which records every interruption of the light beam caused by a person or object passing across the beam. So photocells help count the persons entering an auditorium, provided they enter the hall one by one. They are used for detection of traffic law defaulters: an alarm may be sounded whenever a beam of *(invisible)* radiation is intercepted.

In burglar alarm, (invisible) ultraviolet light is continuously made to fall on a photocell installed at the doorway. A person entering the door interrupts the beam falling on the photocell. The abrupt change in photocurrent is used to start an electric bell ringing. In fire alarm, a number of photocells are installed at suitable places in a building. In the event of breaking out of fire, light radiations fall upon the photocell. This completes the electric circuit through an electric bell or a siren which starts operating as a warning signal.

Photocells are used in the reproduction of sound in motion pictures and in the television camera for scanning and telecasting scenes. They are used in industries for detecting minor flaws or holes in metal sheets.

12.8 Wave Nature of Matter

The dual (wave-particle) nature of light (electromagnetic radiation, in general) comes out clearly from what we have learnt in this and the preceding chapters. The wave nature of light shows up in the phenomena of interference, diffraction and polarisation. On the other hand, in photoelectric effect and Compton effect which involve energy and momentum transfer, radiation behaves as if it is made up of a bunch of particles - the photons. Whether a particle or wave description is best suited for understanding an experiment depends on the nature of the experiment. For example, in the familiar phenomenon of seeing an object by our eye, both descriptions are important. The gathering and focussing mechanism of light by the eye-lens is well described in the wave picture. But its absorption by the rods and cones (of the retina) requires the photon picture of light.

A natural question arises: If radiation has a dual (wave-particle) nature, might not the particles of nature (the electrons, protons, etc.) also exhibit wave-like character? In 1924, the French physicist Louis Victor de Broglie (pronounced as de Broy) (1892-1987) put forward the bold hypothesis that moving particles of matter should display wave-like properties under suitable conditions. He reasoned that nature was symmetrical and that the two basic physical entities - matter and energy, must have symmetrical character. If radiation shows dual aspects, so should matter. De Broglie proposed that the wave length A associated with a particle of momentum p is given as

$$\lambda = \frac{h}{p} = \frac{h}{mv} \qquad (12.5)$$

where *m* is the mass of the particle and *v* its speed. Equation (12.5) is known as the *de Broglie relation* and the wavelength A of the *matter wave* is called *de Broglie wavelength.* The dual aspect of matter is evident in the de Broglie relation. On the left hand side of Eq. (12.5), A, is the attribute of a wave while on the right hand side the momentum p is a typical attribute of a particle. Planck's constant *h* relates the two attributes.

Equation (12.5) for a material particle is basically a hypothesis whose validity can be tested only by experiment. However, it is interesting to see that it is satisfied also by a photon. For a photon, as we have seen,

$$p = h\nu/c \qquad (12.6)$$

Therefore,

$$\frac{h}{p} = \frac{c}{\nu} = \lambda \qquad (12.7)$$

That is, the de Broglie wavelength of a photon given by Eq. (12.5) equals the wavelength of electromagnetic radiation of which the photon is a quantum of energy and momentum.

Clearly, from Eq. (12.5), A is smaller for a heavier particle (large m) or more energetic particle (large *v).* For example, the de Broglie wavelength of a ball of mass 0.12 kg moving with a speed of 20 m s^{-1} is easily calculated :

$$p = mv = 0.12 \text{ kg} \times 20 \text{ m s}^{1} = 2.40 \text{ kg m s}^{-1}$$

$$\lambda = \frac{h}{p} = \frac{6.63\times10^{-34} \text{ J s}}{2.40 \text{ kg m s}^{-1}} = 2.76 \times 10^{-34} \text{ m}$$

Louis Victor de Broglie (1892-1987) French physicist who put forth revolutionary idea of wave nature of matter. This idea was developed by Erwin Schrodinger into a full- fledged theory of quantum mechanics commonly known as wave mechanics. In 1929, he was awarded the Nobel Prize in Physics for his discovery of the wave nature of electrons.

This wavelength Is so small that It Is beyond any measurement. This Is the reason why macroscopic objects In our dally life do not show wave-like properties. On the other hand, In the sub-atomic domain, the wave character of particles is significant and measurable.

Consider an electron (mass m, charge *e)* accelerated from rest through a potential *V*. The kinetic energy *K* of the electron equals the work done (eV) on it by the electric field :

$$K = eV \tag{12.8}$$

Now, $K = \frac{1}{2} mv^2 = \frac{p^2}{2m}$, so that

$$p = \sqrt{2mK} = \sqrt{2meV} \tag{12.9}$$

The de Broglie wavelength A of the electron is then

$$\lambda = \frac{h}{p} = \frac{h}{\sqrt{2mK}} = \frac{h}{\sqrt{2meV}} \tag{12.10}$$

Substituting the numerical values of *h*, *m*, *e*, we get

$$\lambda = \frac{1.227}{\sqrt{V}} \text{ nm} \tag{12.11}$$

where *V* is the magnitude of accelerating potential in volts. For a 120 V accelerating potential, Eq. (12.11) gives λ = 0.112 nm. This wavelength is of the same order as the spacing between the atomic planes in crystals. This suggests that matter waves associated with an electron could be verified by crystal diffraction experiments analogous to X-ray diffraction. We describe the experimental verification of the de Broglie hypothesis in the next section. In 1929, de Broglie was awarded the Nobel Prize in Physics for his discovery of the wave nature of electrons.

The matter-wave picture elegantly incorporated the Heisenberg's *uncertainty principle.* According to the principle, it is not possible to measure *both* the position and momentum of an electron (or any other particle) *at the same time* exactly. There is always some uncertainty (Δx) in the specification of position and some uncertainty (Δp) in the specification of momentum. The product of Δx and Δp is of the order of h* (with $h = h/2\pi$), i.e.,

---

A more rigorous treatment gives $\Delta x\, \Delta p \geq h/2$.

$$\Delta x\, \Delta p = h \tag{12.12}$$

Equation (12.12) allows the possibility that Δx is zero; but then Δp must be infinite in order that the product is non-zero. Similarly, if Δp is zero, Δx must be infinite. Ordinarily, both Δx and Δp are non-zero such that their product is of the order of h.

Now, if an electron has a definite momentum p, (i.e. $\Delta p = 0$), by the de Broglie relation, it has a definite wavelength λ. A wave of definite (single) wavelength extends all over space. By Born's probability interpretation this means that the electron is not localised in any finite region of space. That is, its position uncertainty is infinite ($\Delta x \to \infty$), which is consistent with the uncertainty principle.

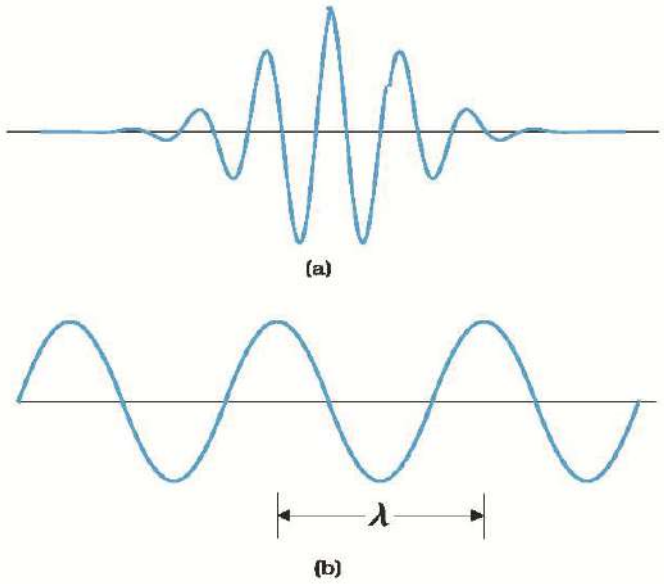


Fig. 12.6 (a) The wave packet description of an electron. The wave packet corresponds to a spread of wavelength around some central wavelength (and hence by de Broglie relation, a spread in momentum). Consequently, It is associated with an uncertainty in position (Δx) and an uncertainty in momentum (Δp). (b) The matter wave corresponding to a definite momentum of an electron extends all over space. In this case, $\Delta p = 0$ and $\Delta x \to \infty$

In general, the matter wave associated with the electron is not extended all over space. It is a wave packet extending over some finite region of space. In that case Axis not infinite but has some finite value depending on the extension of the wave packet. Also, you must appreciate that a wave packet of finite extension does not have a single wavelength. It is built up of wavelengths spread around some central wavelength.

By de Broglie's relation, then, the momentum of the electron will also have a spread - an uncertainty Ap. This is as expected from the uncertainty principle. It can be shown that the wave packet description together with de Broglie relation and Born's probability interpretation reproduce the Heisenberg's uncertainty principle exactly.

In Chapter 13, the de Broglie relation will be seen to justify Bohr's postulate on quantisation of angular momentum of electron in an atom.

Figure 13.6 shows a schematic diagram of (a) a localised wave packet, and (b) an extended wave with fixed wavelength.

Probability Interpretation to Matter Waves

It is worth pausing here to reflect on just what a matter wave associated with a particle, say, an electron, means. Actually, a truly satisfactory physical understanding of the dual nature of matter and radiation has not emerged so far. The great founders of quantum mechanics (Niels Bohr, Albert Einstein, and many others) struggled with this and related concepts for long. Still the deep physical interpretation of quantum mechanics continues to be an area of active research. Despite this, the concept of matter wave has been mathematically introduced in modern quantum mechanics with great success. An important milestone in this connection was when Max Born (1882-1970) suggested a probability interpretation to the matter wave amplitude. According to this, the intensity (square of the amplitude) of the matter wave at a point determines the probability density of the particle at that point. Probability density means probability per unit volume. Thus, if A is the amplitude of the wave at a point, $|A|^2$ AV is the probability of the particle being found in a small volume AV

around that point. Thus, if the intensity of matter wave is large in a certain region, there is a greater probability of the particle being found there than where the intensity is small.

12.9 Davisson and Germer Experiment

The wave nature of electrons was first experimentally verified by C.J. Davisson and L.H. Germer in 1927 and independently by G.P. Thomson, in 1928, who observed diffraction effects with beams of electrons scattered by crystals. Davisson and Thomson shared the Nobel Prize in 1937 for their experimental discovery of diffraction of electrons by crystals.

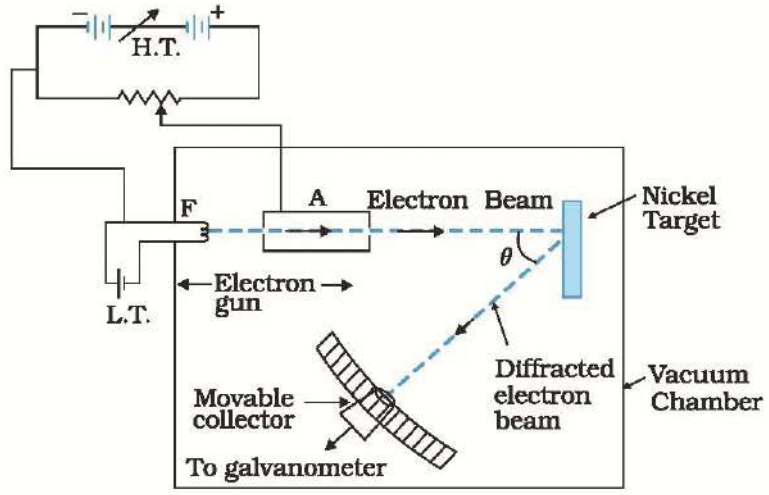


Fig. 12.7 Davisson-Germer electron diffraction arrangement.

The experimental arrange-ment used by Davisson and Germer is schematically shown in Fig. 12.7. It consists of an electron gun which comprises of a tungsten filament F, coated with barium oxide and heated by a low voltage power supply (L.T. or battery). Electrons emitted by the filament are accelerated to a desired velocity by applying suitable potential/voltage from a high voltage power supply (H.T. or battery). They are made to pass through a cylinder with fine holes along its axis, producing a fine collimated beam. The beam is made to fall on the surface of a nickel crystal. The electrons are scattered in all directions by the atoms of the crystal. The intensity of the electron beam, scattered in a given direction, is measured by the electron detector (collector). The detector can be moved on a circular scale and is connected to a sensitive galvanometer, which records the current. The deflection of the galvanometer is proportional to the intensity of the electron beam entering the collector. The apparatus is enclosed in an evacuated chamber. By moving the detector on the circular scale at different positions, the intensity of the scattered electron beam is measured for different values of angle of scattering θ which is the angle between the incident and the scattered electron beams. The variation of the intensity (I) of the scattered electrons with the angle of scattering 0is obtained for different accelerating a voltages.

The experiment was performed by varying the accelerating voltage from 44 V to 68 V. It was noticed that a strong peak appeared in the intensity (I) of the scattered electron for an accelerating voltage of 54V at a scattering angle θ = 50°.

The appearance of the peak in a particular direction is due to the constructive interference of electrons scattered from different layers of the regularly spaced atoms of the crystals. From the electron diffraction measurements, the wavelength of matter waves was found to be 0.165 nm.

The de Broglie wavelength A associated with electrons, using Eq. (12.11), for V= 54 V is given by

$$\lambda = h/p = \frac{1.227}{\sqrt{V}} \text{ nm}$$

$$\lambda = \frac{1.227}{\sqrt{54}} \text{ nm} = 0.167 \text{ nm}$$

Thus, there is an excellent agreement between the theoretical value and the experimentally obtained value of de Broglie wavelength. Davisson- Germer experiment thus strikingly confirms the wave nature of electrons and the de Broglie relation. More recently, in 1989, the wave nature of a beam of electrons was experimentally demonstrated in a double-slit experiment, similar to that used for the wave nature of light. Also, in an experiment in 1994, interference fringes were obtained with the beams of iodine molecules, which are about a million times more massive than electrons.

The de Broglie hypothesis has been basic to the development of modern quantum mechanics. It has also led to the field of electron optics. The wave properties of electrons have been utilised in the design of electron microscope which is a great improvement, with higher resolution, over the optical microscope.

Summary

1. The minimum energy needed by an electron to come out from a metal surface is called the work function of the metal. Energy (greater than the work function (θ_0) required for electron emission from the metal surface can be supplied by suitably heating or applying strong electric field or irradiating it by light of suitable frequency.
2. Photoelectric effect is the phenomenon of emission of electrons by metals when illuminated by light of suitable frequency. Certain metals respond to ultraviolet light while others are sensitive even to the visible light. Photoelectric effect involves conversion of light energy into electrical energy. It follows the law of conservation of energy. The photoelectric emission is an instantaneous process and possesses certain special features.
3. Photoelectric current depends on (i) the intensity of incident light, (ii) the potential difference applied between the two electrodes, and (iii) the nature of the emitter material.
4. The stopping potential (V_0) depends on (i) the frequency of incident light, and (ii) the nature of the emitter material. For a given frequency of incident light, it is independent of its intensity. The stopping potential is directly related to the maximum kinetic energy of electrons emitted : $eV = (1/2)\, m\, v^2_{max} = K_{max}$.
5. Below a certain frequency (threshold frequency) v_0, characteristic of the metal, no photoelectric emission takes place, no matter how large the Intensity may be.
6. The classical wave theory could not explain the main features of photoelectric effect. Its picture of continuous absorption of energy from radiation could not explain the independence of K_{max} on intensity, the existence of v_0 and the instantaneous nature of the process. Einstein explained these features on the basis of photon picture of light. According to this, light is composed of discrete packets of energy called quanta or photons. Each photon carries an energy E (= $h\nu$) and momentum p (= h/λ), which depend on the frequency (v) of incident light and not on its intensity. Photoelectric emission from the metal surface occurs due to absorption of a photon by an electron.
7. Einstein's photoelectric equation is in accordance with the energy conservation law as applied to the photon absorption by an electron in the metal. The maximum kinetic

energy (l/2)m v^2_{max} is equal to the photon energy (hv) minus the work function ϕ_0 (= hv_0) of the target metal:

$$\frac{1}{2} mv^2_{max} = v_0 e = hv - \phi_0 = h(v - v_0)$$

This photoelectric equation explains all the features of the photoelectric effect. Millikan's first precise measurements confirmed the Einstein's photoelectric equation and obtained an accurate value of Planck's constant *h*. This led to the acceptance of particle or photon description (nature) of electromagnetic radiation, introduced by Einstein.

8. Radiation has dual nature: wave and particle. The nature of experiment determines whether a wave or particle description is best suited for understanding the experimental result. Reasoning that radiation and matter should be symmetrical in nature, Louis Victor de Broglie attributed a wave-like character to matter (material particles). The waves associated with the moving material particles are called matter waves or de Broglie waves.
9. The de Broglie wavelength (A) associated with a moving particle is related to its momentum *p* as: $\lambda = h/p$. The dualism of matter is inherent in the de Broglie relation which contains a wave concept (A) and a particle concept (p). The de Broglie wavelength is independent of the charge and nature of the material particle. It is significantly measurable (of the order of the atomic-planes spacing in crystals) only in case of sub-atomic particles like electrons, protons, etc. (due to smallness of their masses and hence, momenta). However, it is indeed very small, quite beyond measurement, in case of macroscopic objects, commonly encountered in everyday life.
10. Electron diffraction experiments by Davisson and Germer, and by G. P. Thomson, as well as many later experiments, have verified and confirmed the wave-nature of electrons. The de Broglie hypothesis of matter waves supports the Bohr's concept of stationary orbits.

Physical Quantity	**Symbol**	**Dimensions**	**Unit**	**Remarks**
Planck's constant	h	$[ML^2T^{-1}]$	J s	$E = hv$
Stopping potential	v_0	$[ML^2T^{-3}A']$	V	$e V_0 = K_{max}$
Work function	ϕ_0	$[ML^2T\sim^2]$	J; eV	$K_{max} = E - \phi_0$
Threshold frequency	v_0	[T"¹]	Hz	$v_0 = \phi_0/h$
de Broglie wavelength	λ	[L]	m	$\lambda = h/p$

VERY SHORT ANSWER QUESTIONS (2 Marks)

1. What are “cathode rays”?
2. What important fact did Millikan’s experiment establish?
3. What is “work function”?
4. What is “photoelectric effect”?
5. Give examples of “photosensitive substances”. Why are they called so?
6. Write down Eienstein’s photoelectric equation.
7. Write down de-Broglie’s relation and explain the terms therein.
8. State Heisenberg’s Uncertainly Principle.

SHORT ANSWER QUESTIONS (4 Marks)

1. What is the effect of (i) intensity of light (ii) potential on photoelectric current?
2. Summarise the photon picture of electro-magnetic radiation.

LONG ANSWER QUESTIONS (8 Marks)

1. How did Einstein’s photo-electric equation explain the effect of intensity and potential on photo-electric current? How did this equation account for the effect of frequency of incident light on stopping potential?

CHAPTER 13

ATOMS

13.1 INTRODUCTION

By the nineteenth century, enough evidence had accumulated in favour of atomic hypothesis of matter. In 1897, the experiments on electric discharge through gases carried out by the English physicist J. J. Thomson (1856–1940) revealed that atoms of different elements contain negatively charged constituents (electrons) that are identical for all atoms. However, atoms on a whole are electrically neutral. Therefore, an atom must also contain some positive charge to neutralise the negative charge of the electrons. But what is the arrangement of the positive charge and the electrons inside the atom? In other words, what is the structure of an atom?

The first model of atom was proposed by J. J. Thomson in 1898. According to this model, the positive charge of the atom is uniformly distributed throughout the volume of the atom and the negatively charged electrons are embedded in it like seeds in a watermelon. This model was picturesquely called *plum pudding model* of the atom. However subsequent studies on atoms, as described in this chapter, showed that the distribution of the electrons and positive charges are very different from that proposed in this model.

We know that condensed matter (solids and liquids) and dense gases at all temperatures emit electromagnetic radiation in which a continuous distribution of several wavelengths is present, though with different intensities. This radiation is considered to be due to oscillations of atoms and molecules, governed by the interaction of each atom or molecule with its neighbours. *In contrast*, light emitted from rarefied gases heated in a flame, or excited electrically in a glow tube such as the familiar neon sign or mercury vapour light has only certain discrete wavelengths. The spectrum appears as a series of bright lines. In such gases, the average spacing between atoms is large. Hence, the radiation emitted can be considered due to individual atoms rather than because of interactions between atoms or molecules.

In the early nineteenth century it was also established that each element is associated with a characteristic spectrum of radiation, for example, hydrogen always gives a set of lines with fixed relative position between the lines. This fact suggested an intimate relationship between the internal structure of an atom and the spectrum of radiation emitted by it. In 1885, Johann Jakob Balmer (1825–1898) obtained a simple empirical formula which gave the wavelengths of a group of lines emitted by atomic hydrogen. Since hydrogen is simplest of the elements known, we shall consider its spectrum in detail in this chapter.

Ernst Rutherford (1871–1937), a former research student of J. J. Thomson, was engaged in experiments on α-particles emitted by some radioactive elements. In 1906, he proposed a classic experiment of scattering of these α-particles by atoms to investigate the atomic structure. This experiment was later performed around 1911 by Hans Geiger (1882–1945) and Ernst Marsden (1889–1970, who was 20 year-old student and had not yet earned his bachelor's degree). The details are discussed in Section 13.2. The explanation of the results led to the birth of Rutherford's planetary model of atom (also called the *nuclear model of the atom*). According to this the entire positive charge and most of the mass of the atom is concentrated in a small volume called the nucleus with electrons revolving around the nucleus just as planets revolve around the sun.

Rutherford's nuclear model was a major step towards how we see the atom today. However, it could not explain why atoms emit light of only discrete wavelengths. How could an atom as simple as hydrogen, consisting of a single electron and a single proton, emit a complex spectrum of specific wavelengths? In the classical picture of an atom, the electron revolves round the nucleus much like the way a planet revolves round the sun. However, we shall see that there are some serious difficulties in accepting such a model.

ERNST RUTHERFORD (1871 – 1937)

Ernst Rutherford (1871 – 1937) New Zealand born, British physicist who did pioneering work on radioactive radiation. He discovered alpha-rays and beta-rays. Along with Federick Soddy, he created the modern theory of radioactivity. He studied the 'emanation' of thorium and discovered a new noble gas, an isotope of radon, now known as thoron. By scattering alpha-rays from the metal foils, he discovered the atomic nucleus and proposed the planetary model of the atom. He also estimated the approximate size of the nucleus.

13.2 ALPHA-PARTICLE SCATTERING AND RUTHERFORD'S NUCLEAR MODEL OF ATOM

At the suggestion of Ernst Rutherford, in 1911, H. Geiger and E. Marsden performed some experiments. In one of their experiments, as shown in Fig. 13.1, they directed a beam of 5.5 MeV α-particles emitted from a $_{83}Bi^{214}$ radioactive source at a thin metal foil made of gold. Figure 13.2 shows a schematic diagram of this experiment. Alpha-particles emitted by a $_{83}Bi^{214}$ radioactive source were collimated into a narrow beam by their passage through lead bricks. The beam was allowed to fall on a thin foil of gold of thickness 2.1×10^{-7} m. The scattered alpha-particles were observed through a rotatable detector consisting of zinc sulphide screen and a microscope. The scattered alpha-particles on striking the screen produced brief light flashes or scintillations. These flashes may be viewed through a microscope and the distribution of the number of scattered particles may be studied as a function of angle of scattering.

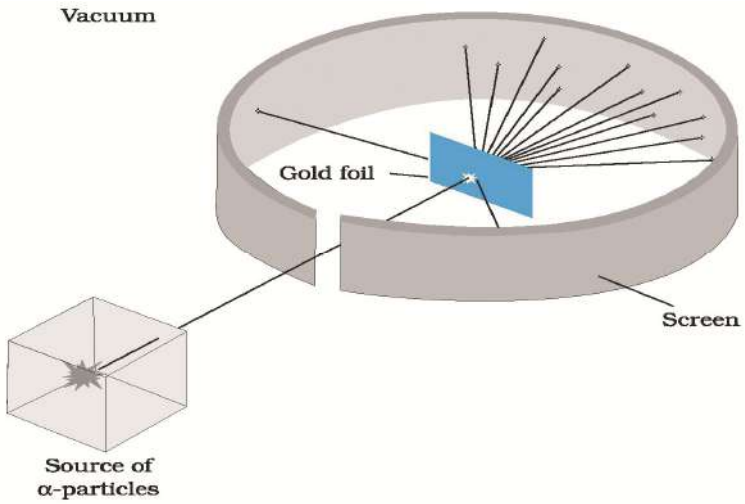


Fig. 13.1 Schematic arrangement of the Geiger-Marsden experiment. The entire apparatus is placed in a vaccum chamber (not shown in this figure)

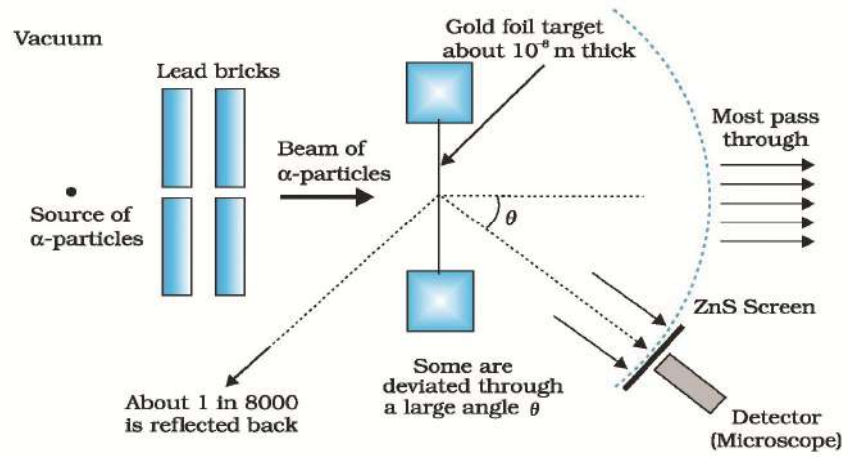


Fig. 13.2 Schematic arrangement of the Geiger-Marsden experiment.

A typical graph of the total number of α-particles scattered at different angles, in a given interval of time, is shown in Fig. 13.3. The dots in this figure represent the data points and the solid curve is the theoretical prediction based on the assumption that the target atom has a small, dense, positively charged nucleus. Many of the α-particles pass through the foil. It means that they do not suffer any collisions. Only about 0.14% of the incident α-particles scatter by more than 1°; and about 1 in 8000 deflect by more than 90°. Rutherford argued that, to deflect the α-particle backwards, it must experience a large repulsive force. This force could be provided if the greater part of the mass of the atom and its positive charge were concentrated tightly at its centre. Then the incoming α-particle could get very close to the positive charge without penetrating it, and such a close encounter would result in a large deflection. This agreement supported the hypothesis of the nuclear atom. This is why Rutherford is credited with the *discovery* of the nucleus.

In Rutherford's nuclear model of the atom, the entire positive charge and most of the mass of the atom are concentrated in the nucleus with the electrons some distance away. The electrons would be moving in orbits about the nucleus just as the planets do around the sun. Rutherford's experiments suggested the size of the nucleus to be about 10^{-15} m to 10^{-14} m. From kinetic theory, the size of an atom was known to be 10^{-10} m, about 10,000 to 100,000 times larger than the size of the nucleus. Thus, the electrons would seem to be at a distance from the nucleus of about 10,000 to 100,000 times the size of the nucleus itself. Thus, most of an atom is empty space. With the atom being largely empty space, it is easy to see why most α-particles go right through a thin metal foil. However, when α-particle happens to come near a nucleus, the intense electric field there scatters it through a large angle. The atomic electrons, being so light, do not appreciably affect the α-particles.

The scattering data shown in Fig. 13.3 can be analysed by employing Rutherford's nuclear model of the atom. As the gold foil is very thin, it can be assumed that α-particles will suffer not more than one scattering during their passage through it. Therefore, computation of the trajectory of an alpha-particle scattered by a single nucleus is enough. Alpha- particles are nuclei of helium atoms and, therefore, carry two units, 2*e*, of positive charge and have the mass of the helium atom.

The charge of the gold nucleus is Ze, where Z is the atomic number of the atom; for gold $Z = 79$. Since the nucleus of gold is about 50 times heavier than an α-particle, it is reasonable to assume that it remains stationary throughout the scattering process. Under these assumptions, the trajectory of an alpha-particle can be computed employing Newton's second law of motion and the Coulomb's law for electrostatic force of repulsion between the alpha-particle and the positively charged nucleus.

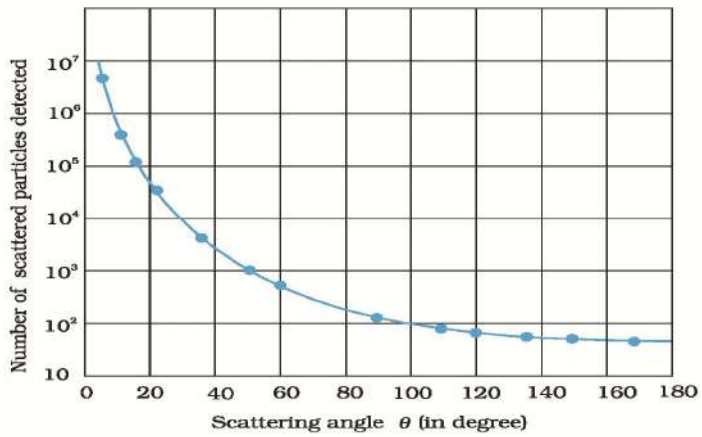


Fig.13.3 Experimental data points (shown by dots) on scattering of α-particles by a thin foil at different angles obtained by Geiger and Marsden using the setup shown in Figs. 13.1 and 13.2. Rutherford's nuclear model predicts the solid curve which is seen to be in good agreement with experiment.

The magnitude of this force is

$$F = \frac{1}{4\pi\varepsilon_0}\frac{(2e)(Ze)}{r^2} \qquad (13.1)$$

where r is the distance between the α-particle and the nucleus. The force is directed along the line joining the α-particle and the nucleus. The magnitude and direction of the force on an α-particle continuously changes as it approaches the nucleus and recedes away from it.

13.2.1 Alpha-particle trajectory

The trajectory traced by an α-particle depends on the impact parameter, b of collision. The *impact parameter* is the perpendicular distance of the initial velocity vector of the α-particle from the centre of the nucleus (Fig. 13.4). A given beam of α-particles has a distribution of impact parameters b, so that the beam is scattered in various directions with different probabilities (Fig. 13.4). (In a beam, all particles have nearly same kinetic energy.) It is seen that an α-particle close to the nucleus (small impact parameter) suffers large scattering. In case of head-on collision, the impact parameter is minimum and the α-particle rebounds back ($\theta \cong \pi$). For a large impact parameter, the α-particle goes nearly undeviated and has a small deflection ($\theta \cong 0$).

The fact that only a small fraction of the number of incident particles rebound back indicates that the number of α-particles undergoing head on collision is small. This, in turn, implies that the mass of the atom is concentrated in a small volume. Rutherford scattering therefore, is a powerful way to determine an upper limit to the size of the nucleus.

13.2.2 Electron orbits

The Rutherford nuclear model of the atom which involves classical concepts, pictures the atom as an electrically neutral sphere consisting of a very small, massive and positively charged nucleus at the centre surrounded by the revolving electrons in their respective dynamically stable orbits. The electrostatic force of attraction, F_e between the revolving electrons and the nucleus provides the requisite centripetal force (F_c) to keep them in their orbits. Thus, for a dynamically stable orbit in a hydrogen atom

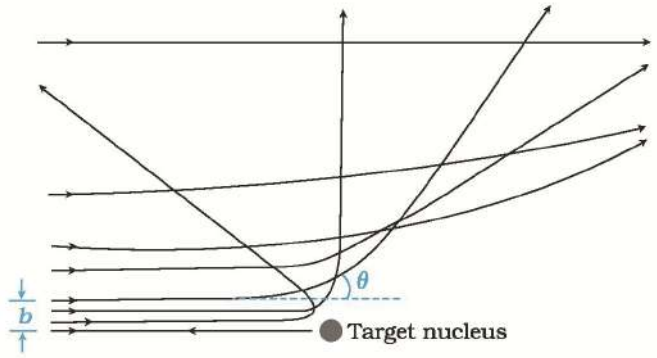


Fig.13.4 Trajectory of α-particles in the coulomb field of a target nucleus. The impact parameter, *b* and scattering angle θ are also depicted.

$$F_e = F_c$$

$$\frac{mv^2}{r} = \frac{1}{4\pi\varepsilon_0}\frac{e^2}{r^2} \tag{13.2}$$

Thus the relation between the orbit radius and the electron velocity is

$$r = \frac{e^2}{4\pi\varepsilon_0 mv^2} \tag{13.3}$$

The kinetic energy (K) and electrostatic potential energy (U) of the electron in hydrogen atom are

$$K = \frac{1}{2}mv^2 = \frac{e^2}{8\pi\varepsilon_0 r} \text{ and } U = -\frac{e^2}{4\pi\varepsilon_0 r}$$

(The negative sign in U signifies that the electrostatic force is in the $-r$ direction.) Thus the total energy E of the electron in a hydrogen atom is

$$E = K + U = \frac{e^2}{8\pi\varepsilon_0 r} - \frac{e^2}{4\pi\varepsilon_0 r}$$

$$= \frac{e^2}{8\pi\varepsilon_0 r} \tag{13.4}$$

The total energy of the electron is negative. This implies the fact that the electron is bound to the nucleus. If E were positive, an electron will not follow a closed orbit around the nucleus.

13.3 ATOMIC SPECTRA

As mentioned in Section 13.1, each element has a characteristic spectrum of radiation, which it emits. When an atomic gas or vapour is excited at low pressure, usually by passing an electric current through it, the emitted radiation has a spectrum which contains certain specific wavelengths only. A spectrum of this kind is termed as emission line spectrum and it consists of bright lines on a dark background. The spectrum emitted by atomic hydrogen is shown in Fig.

13.5. Study of emission line spectra of a material can therefore serve as a type of "fingerprint" for identification of the gas. When white light passes through a gas and we analyse the transmitted light using a spectrometer we find some dark lines in the spectrum. These dark lines correspond precisely to those wavelengths which were found in the emission line spectrum of the gas. This is called the *absorption spectrum* of the material of the gas.

13.3.1 Spectral series

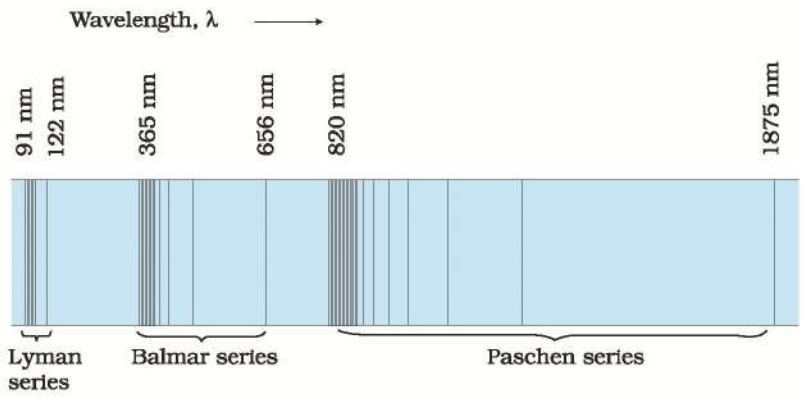


Fig. 13.5 Emission lines in the spectrum of hydrogen.

We might expect that the frequencies of the light emitted by a particular element would exhibit some regular pattern. Hydrogen is the simplest atom and therefore, has the simplest spectrum. In the observed spectrum, however, at first sight, there does not seem to be any resemblance of order or regularity in spectral lines. But the spacing between lines within certain sets of the hydrogen spectrum decreases in a regular way (Fig. 13.5). Each of these sets is called a *spectral series*. In 1885, the first such series was observed by a Swedish school teacher Johann Jakob Balmer (1825–1898) in the visible region of the hydrogen spectrum. This series is called *Balmer series* (Fig. 13.6). The line with the longest wavelength, 656.3 nm in the red is called Hα; the next line with wavelength 486.1 nm in the blue- green is called Hβ, the third line 434.1 nm in the violet is called Hγ; and so on. As the wavelength decreases, the lines appear closer together and are weaker in intensity.

Balmer found a simple empirical formula for the observed wavelengths

$$\frac{1}{\lambda} = R\left(\frac{1}{2^2} - \frac{1}{n^2}\right) \qquad (13.5)$$

where $\lambda$is the wavelength, R is a constant called the *Rydberg constant*, and n may have integral values 3, 4, 5, etc. The value of R is 1.097×10^7 m^{-1}. This equation is also called Balmer formula.

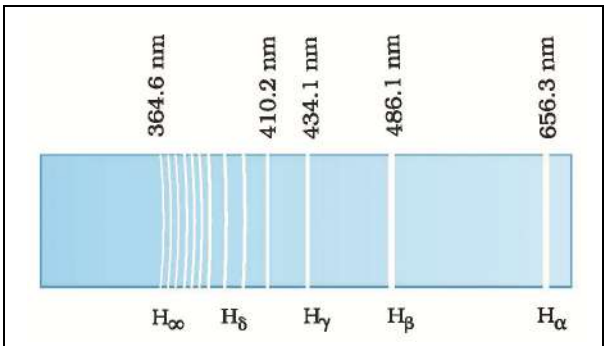


Fig. 13.6 Balmer series in the emission spectrum of hydrogen.

Taking n = 3 in Eq. (13.5), one obtains the wavelength of the Hα line:

$$\frac{1}{\lambda} = 1.097 \times 10^7\left(\frac{1}{2^2} - \frac{1}{3^2}\right)\text{m}^{-1}$$

$$= 1.522 \times 10^6 \text{ m}^{-1}$$

i.e., λ = 656.3 nm

For n = 4, one obtains the wavelength of Hβ line, etc. For $n = \infty$, one obtains the limit of the series, at λ = 364.6 nm. This is the shortest wavelength in the Balmer series. Beyond this limit, no further distinct lines appear, instead only a faint continuous spectrum is seen.

Other series of spectra for hydrogen were subsequently discovered. These are known, after their discoverers, as Lyman, Paschen, Brackett, and Pfund series. These are represented by the formulae:

Lyman series :

$$\frac{1}{\lambda} = R\left(\frac{1}{1^2} - \frac{1}{n^2}\right) \qquad n = 2, 3, 4..... \qquad (13.6)$$

Paschen series :

$$\frac{1}{\lambda} = R\left(\frac{1}{3^2} - \frac{1}{n^2}\right) \qquad n = 4,5,6... \qquad (13.7)$$

Brackett series :

$$\frac{1}{\lambda} = R\left(\frac{1}{4^2} - \frac{1}{n^2}\right) \qquad n = 5,6,7... \qquad (13.8)$$

Pfund series :

$$\frac{1}{\lambda} = R\left(\frac{1}{5^2} - \frac{1}{n^2}\right) \qquad n = 6,7,8... \qquad (13.9)$$

The Lyman series is in the ultraviolet, and the Paschen and Brackett series are in the infrared region.

The Balmer formula Eq. (13.5) may be written in terms of frequency of the light, recalling that

$$c = \nu\lambda$$

or $$\frac{1}{\lambda} = \frac{\nu}{c}$$

Thus, Eq. (13.5) becomes

$$\nu = Rc\left(\frac{1}{2^2} - \frac{1}{n^2}\right) \qquad (13.10)$$

There are only a few elements (hydrogen, singly ionised helium, and doubly ionised lithium) whose spectra can be represented by simple formula like Eqs. (13.5) – (13.9).

Equations (13.5) – (13.9) are useful as they give the wavelengths that hydrogen atoms radiate or absorb. However, these results are empirical and do not give any reasoning why only certain frequencies are observed in the hydrogen spectrum.

13.4 BOHR MODEL OF THE HYDROGEN ATOM

The model of the atom proposed by Rutherford assumes that the atom, consisting of a central nucleus and revolving electron is stable much like sun-planet system which the model imitates. However, there are some fundamental differences between the two situations. While the planetary system is held by gravitational force, the nucleus-electron system being charged objects, interact by Coulomb's Law of force. We know that an object which moves in a circle is being constantly accelerated – the acceleration being centripetal in nature. According to classical

electromagnetic theory, an accelerating charged particle emits radiation in the form of electromagnetic waves. The energy of an accelerating electron should therefore, continuously decrease. The electron would spiral inward and eventually fall into the nucleus (Fig. 13.7). Thus, such an atom can not be stable. Further, according to the classical electromagnetic theory, the frequency of the electromagnetic waves emitted by the revolving electrons is equal to the frequency of revolution. As the electrons spiral inwards, their angular velocities and hence their frequencies would change continuously, and so will the frequency of the light emitted. Thus, they would emit a continuous spectrum, in contradiction to the line spectrum actually observed. Clearly Rutherford model tells only a part of the story implying that the classical ideas are not sufficient to explain the atomic structure.

NIELS HENRIK DAVID BOHR (1885 – 1962)

Niels Henrik David Bohr (1885 – 1962) Danish physicist who explained the spectrum of hydrogen atom based on quantum ideas. He gave a theory of nuclear fission based on the liquid- drop model of nucleus. Bohr contributed to the clarification of conceptual problems in quantum mechanics, in particular by proposing the comple- mentary principle.

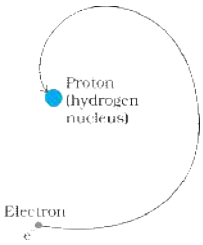


Fig. 13.7 An accelerated atomic electron must spiral into the nucleus as it loses energy.

It was Niels Bohr (1885 – 1962) who made certain modifications in this model by adding the ideas of the newly developing quantum hypothesis. Niels Bohr studied in Rutherford's laboratory for several months in 1912 and he was convinced about the validity of Rutherford nuclear model. Faced with the dilemma as discussed above, Bohr, in 1913, concluded that in spite of the success of electromagnetic theory in explaining large-scale phenomena, it could not be applied to the processes at the atomic scale. It became clear that a fairly radical departure from the established principles of classical mechanics and electromagnetism would be needed to understand the structure of atoms and the relation of atomic structure to atomic spectra. Bohr combined classical and early quantum concepts and gave his theory in the form of three postulates. These are :

(i) Bohr's first postulate was that *an electron in an atom could revolve in certain stable orbits without the emission of radiant energy*, contrary to the predictions of electromagnetic theory. According to this postulate, each atom has certain definite stable states in which it can exist, and each possible state has definite total energy. These are called the stationary states of the atom.

(ii) Bohr's second postulate defines these stable orbits. This postulate states that the *electron* revolves around the nucleus *only in those orbits for which the angular momentum is some integral multiple of* $h/2\pi$ where h is the Planck's constant (= 6.6×10^{-34} J s). Thus the angular momentum (L) of the orbiting electron is quantised. That is

$$L = nh/2\pi \qquad (13.11)$$

(iii) Bohr's third postulate incorporated into atomic theory the early quantum concepts that had been developed by Planck and Einstein. It states that *an electron might make a transition from one of its specified non-radiating orbits to another of lower energy. When it does so,*

a photon is emitted having energy equal to the energy difference between the initial and final states. The frequency of the emitted photon is then given by

$$h\nu = Ei - Ef \qquad (13.12)$$

where E_i and E_f are the energies of the initial and final states and $E_i > E_f$.

For a hydrogen atom, Eq. (13.4) gives the expression to determine the energies of different energy states. But then this equation requires the radius r of the electron orbit. To calculate r, Bohr's second postulate about the angular momentum of the electron–the quantisation condition – is used. The angular momentum L is given by

$$L = mvr$$

Bohr's second postulate of quantisation [Eq. (13.11)] says that the allowed values of angular momentum are integral multiples of $h/2\pi$.

$$L_n = mv_n r_n = \frac{nh}{2\pi} \qquad (13.13)$$

where n is an integer, r_n is the radius of n^{th} possible orbit and v_n is the speed of moving electron in the n orbit. The allowed orbits are numbered 1, 2, 3 ..., according to the values of n, which is called the *principal quantum number* of the orbit.

From Eq. (13.3), the relation between v_n and r_n is

$$v_n = \frac{e}{\sqrt{4\pi\varepsilon_0 m r_n}}$$

Combining it with Eq. (13.13), we get the following expressions for v_n and r_n,

$$v_n = \frac{1}{n}\frac{e^2}{4\pi\varepsilon_0}\frac{1}{(h/2\pi)} \qquad (13.14)$$

and

$$r_n = \left(\frac{n^2}{m}\right)\left(\frac{h}{2\pi}\right)^2 \frac{4\pi\varepsilon_0}{e^2} \qquad (13.15)$$

Eq. (13.14) depicts that the orbital speed in the n^{th} orbit falls by a factor of n. Using Eq. (13.15), the size of the innermost orbit (n = 1) can be obtained as

$$r_1 = \frac{h^2\varepsilon_0}{\pi m e^2}$$

This is called the Bohr radius, represented by the symbol a_0. Thus,

$$a_0 = \frac{h^2\varepsilon_0}{\pi m e^2} \qquad (13.16)$$

Substitution of values of h, m, ε_0 and e gives $a_0 = 5.29 \times 10^{-11}$ m. From Eq. (13.15), it can also be seen that the radii of the orbits increase as n^2. The total energy of the electron in the stationary states of the hydrogen atom can be obtained by substituting the value of orbital radius in Eq. (13.4) as

$$E_n = -\left(\frac{e^2}{8\pi\varepsilon_0}\right)\left(\frac{m}{n^2}\right)\left(\frac{2\pi}{h}\right)^2\left(\frac{e^2}{4\pi\varepsilon_0}\right)$$

or $$E_n = -\frac{me^4}{8n^2\varepsilon_0{}^2h^2} \qquad (13.17)$$

Substituting values, Eq. (13.17) yields

$$E_n = -\frac{2.18\times10^{-18}}{n^2}\ \text{J} \qquad (13.18)$$

Atomic energies are often expressed in electron volts (eV) rather than joules. Since 1eV = 1.6×10^{-19} J, Eq. (13.18) can be rewritten as

$$E_n = -\frac{13.6}{n^2}\text{eV} \qquad (13.19)$$

The negative sign of the total energy of an electron moving in an orbit means that the electron is bound with the nucleus. Energy will thus be required to remove the electron from the hydrogen atom to a distance infinitely far away from its nucleus (or proton in hydrogen atom).

The derivation of Eqs. (13.17) – (13.19) involves the assumption that the electronic orbits are circular, though orbits under inverse square force are, in general elliptical. (Planets move in elliptical orbits under the inverse square gravitational force of the sun.) However, it was shown by the German physicist Arnold Sommerfeld (1868 – 1951) that, when the restriction of circular orbit is relaxed, these equations continue to hold even for elliptic orbits.

ORBIT VSSTATE (ORBITAL PICTURE) OF ELECTRON IN ATOM

We are introduced to the Bohr Model of atom one time or the other in the course of physics. This model has its place in the history of quantum mechanics and particularly in explaining the structure of an atom. It has become a milestone since Bohr introduced the revolutionary idea of definite energy orbits for the electrons, contrary to the classical picture requiring an accelerating particle to radiate. Bohr also introduced the idea of quantisation of angular momentum of electrons moving in definite orbits. Thus it was a semi-classical picture of the structure of atom.

Now with the development of quantum mechanics, we have a better understanding of the structure of atom. Solutions of the Schrödinger wave equation assign a wave-like description to the electrons bound in an atom due to attractive forces of the protons.

An orbit of the electron in the Bohr model is the circular path of motion of an electron around the nucleus. But according to quantum mechanics, we cannot associate a definite path with the motion of the electrons in an atom. We can only talk about the probability of finding an electron in a certain region of space around the nucleus. This probability can be inferred from the one-electron wave function called the *orbital*. This function depends only on the coordinates of the electron.

It is therefore essential that we understand the subtle differences that exist in the two models:

- Bohr model is valid for only one-electron atoms/ions; an energy value, assigned to each orbit, depends on the principal quantum number *n* in this model. We know that energy associated with a stationary state of an electron depends on *n* only, for one-electron atoms/ions. For a multi-electron atom/ion, this is not true.

❖ The solution of the Schrödinger wave equation, obtained for hydrogen-like atoms/ ions, called the wave function, gives information about the probability of finding an electron in various regions around the nucleus. This *orbital* has no resemblance whatsoever with the *orbit* defined for an electron in the Bohr model.

13.4.1 Energy levels

The energy of an atom is the *least* (largest negative value) when its electron is revolving in an orbit closest to the nucleus i.e., the one for which $n = 1$. For $n = 2, 3, ...$ the absolute value of the energy E is smaller, hence the energy is progressively larger in the outer orbits. The *lowest* state of the atom, called the *ground state*, is that of the lowest energy, with the electron revolving in the orbit of smallest radius, the Bohr radius, a_0. The energy of this state ($n = 1$), E_1 is –13.6 eV. Therefore, the minimum energy required to

free the electron from the ground state of the hydrogen atom is 13.6 eV. It is called the *ionisation energy* of the hydrogen atom. This prediction of the Bohr's model is in excellent agreement with the experimental value of ionisation energy.

At room temperature, most of the hydrogen atoms are in *ground state*. When a hydrogen atom receives energy by processes such as electron collisions, the atom may acquire sufficient energy to raise the electron to higher energy states. The atom is then said to be in an *excited state*. From Eq. (13.19), for $n = 2$; the energy E_2 is –3.40 eV. It means that the energy required to excite an electron in hydrogen atom to its first excited state, is an energy equal to $E_2 - E_1 = -3.40$ eV – (–13.6) eV = 10.2 eV. Similarly, $E_3 = -1.51$ eV and $E_3 - E_1 = 12.09$ eV, or to excite the hydrogen atom from its ground state ($n = 1$) to second excited state ($n = 3$), 12.09 eV energy is required, and so on. From these excited states the electron can then fall back to a state of lower energy, emitting a photon in the process. Thus, as the excitation of hydrogen atom increases (that is as n increases) the value of minimum energy required to free the electron from the excited atom decreases.

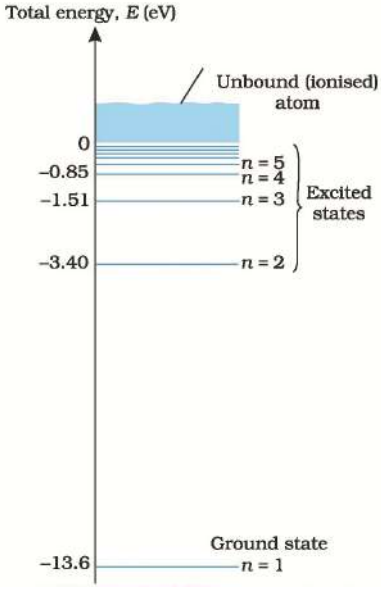


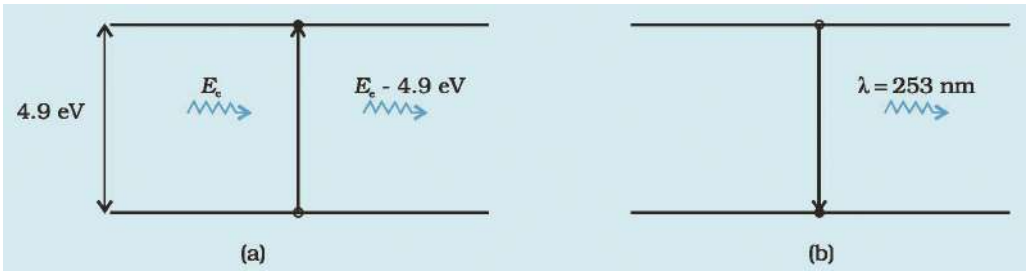
Fig. 13.8 The energy level diagram for the hydrogen atom. The electron in a hydrogen atom at room temperature spends most of its time in the ground state. To ionise a hydrogen atom an electron from the ground state, 13.6 eV of energy must be supplied. (The horizontal lines specify the presence of allowed energy states.)

The energy level diagram***** for the stationary states of a hydrogen atom, computed from Eq. (13.19), is given in Fig. 13.8. The principal quantum number n labels the stationary states in the ascending order of energy. In this diagram, the highest energy state corresponds to $n = \infty$ in Eq, (13.19) and has an energy of 0 eV. This is the energy of the atom when the electron is completely removed ($r = \infty$) from the nucleus and is at rest. Observe how the energies of the excited states come closer and closer together as n increases.

* An electron can have any total energy above $E = 0$ eV. In such situations the electron is free. Thus there is a continuum of energy states above $E = 0$ eV, as shown in Fig. 13.8.

FRANCK – HERTZ EXPERIMENT

The existence of discrete energy levels in an atom was directly verified in 1914 by James Franck and Gustav Hertz. They studied the spectrum of mercury vapour when electrons having different kinetic energies passed through the vapour. The electron energy was varied by subjecting the electrons to electric fields of varying strength. The electrons collide with the mercury atoms and can transfer energy to the mercury atoms. This can only happen when the energy of the electron is higher than the energy difference between an energy level of Hg occupied by an electron and a higher unoccupied level (see Figure). For instance, the difference between an occupied energy level of Hg and a higher unoccupied level is 4.9 eV. If an electron of having an energy of 4.9 eV or more passes through mercury, an electron in mercury atom can absorb energy from the bombarding electron and get excited to the higher level [Fig (a)]. The colliding electron's kinetic energy would reduce by this amount.



The excited electron would subsequently fall back to the ground state by emission of radiation [Fig. (b)]. The wavelength of emitted radiation is:

$$\lambda = \frac{hc}{E} = \frac{6.625\times10^{-34}\times3\times10^{8}}{4.9\times1.6\times10^{-19}} = 253 \text{ nm}$$

By direct measurement, Franck and Hertz found that the emission spectrum of mercury has a line corresponding to this wavelength. For this experimental verification of Bohr's basic ideas of discrete energy levels in atoms and the process of photon emission, Frank and Hertz were awarded the Nobel prize in 1925.

13.5 THE LINE SPECTRA OF THE HYDROGEN ATOM

According to the third postulate of Bohr's model, when an atom makes a transition from the higher energy state with quantum number n_i to the lower energy state with quantum number $n_f (n_f < n_i)$, the difference of energy is carried away by a photon of frequency ν_{if} such that

$$h\nu_{if} = En_i - En_f \qquad (13.20)$$

Using Eq. (13.16), for *Enf* and *Eni*, we get

$$h\nu_{if} = \frac{me^4}{8e_0^2h^2}\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right) \qquad (13.21)$$

$$\text{or } \nu_{if} = \frac{me^4}{8e_0^2h^2}\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right) \qquad (13.22)$$

Equation (13.21) is the Rydberg formula, for the spectrum of the hydrogen atom. In this relation, if we take n_f= 2 and n_i= 3, 4, 5..., it reduces to a form similar to Eq. (13.10) for the Balmer series. The Rydberg
constant R is readily identified to be

$$R = \frac{me^4}{8\varepsilon_0^2 h^3 c} \qquad (13.23)$$

If we insert the values of various constants in Eq. (13.23), we get

$R = 1.03 \times 10^7\ m^{-1}$

This is a value very close to the value ($1.097 \times 10^7\ m^{-1}$) obtained from the empirical Balmer formula. This agreement between the theoretical and experimental values of the Rydberg constant provided a direct and striking confirmation of the Bohr's model.

Since both $n_f$and $n_i$are integers, this immediately shows that in transitions between different atomic levels, light is radiated in various discrete frequencies. For hydrogen spectrum, the Balmer formula corresponds to n_f= 2 and n_i= 3, 4, 5, etc. The results of the Bohr's model suggested the presence of other series spectra for hydrogen atom–those corresponding to transitions resulting from n_f = 1 and n_i= 2, 3, etc.; n_f= 3 and n_i = 4, 5, etc., and so on. Such series were identified in the course of spectroscopic investigations and are known as the L yman, Balmer, Paschen, Brackett, and Pfund series. The electronic transitions corresponding to these series are shown in Fig. 13.9.

The various lines in the atomic spectra are produced when electrons jump from higher energy state to a lower energy state and photons are emitted. These spectral lines are called emission lines. But when an atom absorbs a photon that has precisely the same energy needed by the electron in a lower energy state to make transitions to a higher energy state, the process is called absorption. Thus if photons with a continuous range of frequencies pass through a rarefied gas and then are analysed with a spectrometer, a series of dark spectral absorption lines appear in the continuous spectrum. The dark lines indicate the frequencies that have been absorbed by the atoms of the gas.

The explanation of the hydrogen atom spectrum provided by Bohr's model was a brilliant

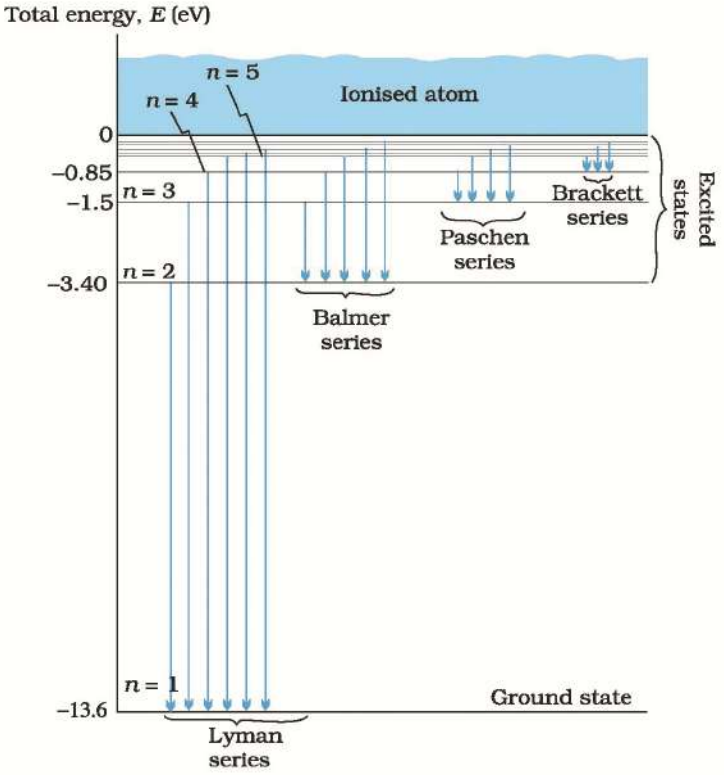


Fig. 13.9 Line spectra originate in transitions between energy levels.

achievement, which greatly stimulated progress towards the modern quantum theory. In 1922, Bohr was awarded Nobel Prize in Physics.

13.6 DE BROGLIE'S EXPLANATION OF BOHR'S SECOND POSTULATE OF QUANTISATION

Of all the postulates, Bohr made in his model of the atom, perhaps the most puzzling is his second postulate. It states that the angular momentum of the electron orbiting around the nucleus is quantised (that is, $L_n = nh/2\pi$; $n = 1, 2, 3 \ldots$). Why should the angular momentum have only those values that are integral multiples of $h/2\pi$? The French physicist Louis de Broglie explained this puzzle in 1923, ten years after Bohr proposed his model.

We studied, in Chapter 12, about the de Broglie's hypothesis that material particles, such as electrons, also have a wave nature. C. J. Davisson and L. H. Germer later experimentally verified the wave nature of electrons in 1927. Louis de Broglie argued that the electron in its circular orbit, as proposed by Bohr, must be seen as a particle wave. In analogy to waves travelling on a string, particle waves too can lead to standing waves under resonant conditions. We know that when a string is plucked, a vast number of wavelengths are excited. However only those wavelengths survive which have nodes at the ends and form the standing wave in the string. It means that in a string, standing waves are formed when the total distance travelled by a wave down the string and back is one wavelength, two wavelengths, or any integral number of wavelengths. Waves with other wavelengths interfere with themselves upon reflection and their amplitudes quickly drop to zero. For an electron moving in n^{th} circular orbit of radius r_n, the total distance is the circumference of the orbit, $2\pi r_n$. Thus

$$2\pi r_n = n\lambda, \quad n = 1, 2, 3... \qquad (13.24)$$

Figure 13.10 illustrates a standing particle wave on a circular orbit for $n = 4$, i.e., $2\pi r_n = 4\lambda$, where λ is the de Broglie wavelength of the electron moving in n orbit. From Chapter 11, we have $\lambda = h/p$, where p is the magnitude of the electron's momentum. If the speed of the electron is much less than the speed of light, the momentum is mv_n. Thus, $\lambda = h/mv_n$. From Eq. (13.24), we have

$$2\pi r_n = n\, h/mv_n \qquad \text{or} \qquad m\, v_n\, r_n = nh/2\pi$$

This is the quantum condition proposed by Bohr for the angular momentum of the electron [Eq. (13.13)]. In Section 12.5, we saw that this equation is the basis of explaining the discrete orbits and energy levels in hydrogen atom. Thus de Broglie hypothesis provided an explanation for Bohr's second postulate for the quantisation of angular momentum of the orbiting electron. The quantised electron orbits and energy states are due to the wave nature of the electron and only resonant standing waves can persist.

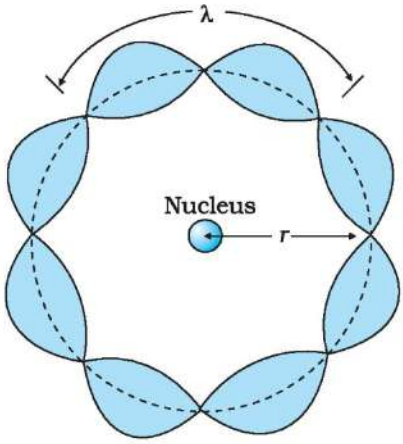


Fig. 13.10 A standing wave is shown on a circular orbit where four de Broglie wavelengths fit into the circumference of the orbit.

Bohr's model, involving classical trajectory picture (planet-like electron orbiting the nucleus), correctly predicts the gross features of the hydrogenic atoms*****, in particular, the

frequencies of the radiation emitted or selectively absorbed. This model however has many limitations. Some are :

(i) The Bohr model is applicable to hydrogenic atoms. It cannot be extended even to mere two electron atoms such as helium. The analysis of atoms with more than one electron was attempted on the lines of Bohr's model for hydrogenic atoms but did not meet with any success. Difficulty lies in the fact that each electron interacts not only with the positively charged nucleus but also with all other electrons.

The formulation of Bohr model involves electrical force between positively charged nucleus and electron. It does not include the electrical forces between electrons which necessarily appear in multi-electron atoms.

(ii) While the Bohr's model correctly predicts the frequencies of the light emitted by hydrogenic atoms, the model is unable to explain the relative intensities of the frequencies in the spectrum. In emission spectrum of hydrogen, some of the visible frequencies have weak intensity, others strong. Why? Experimental observations depict that some transitions are more favoured than others. Bohr's model is unable to account for the intensity variations.

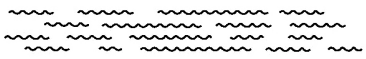
Bohr's model presents an elegant picture of an atom and cannot be generalised to complex atoms. For complex atoms we have to use a new and radical theory based on Quantum Mechanics, which provides a more complete picture of the atomic structure.

***** Hydrogenic atoms are the atoms consisting of a nucleus with positive charge +Ze and a single electron, where Z is the proton number. Examples are hydrogen atom, singly ionised helium, doubly ionised lithium, and so forth. In these atoms more complex electron-electron interactions are nonexistent.

LASER LIGHT

Imagine a crowded market place or a railway platform with people entering a gate and going towards all directions. Their footsteps are random and there is no phase correlation between them. On the other hand, think of a large number of soldiers in a regulated march. Their footsteps are very well correlated. See figure here.

(a) Light from a bulb



(b) Laser light

This is similar to the difference between light emitted by an ordinary source like a candle or a bulb and that emitted by a laser. The acronym LASER stands for Light Amplification by Stimulated Emission of Radiation. Since its development in 1960, it has entered into all areas of science and technology. It has found applications in physics, chemistry, biology, medicine, surgery, engineering, etc. There are low power lasers, with a power of 0.5 mW, called pencil lasers, which serve as pointers. There are also lasers of different power, suitable for delicate surgery of eye or glands in the stomach. Finally, there are lasers which can cut or weld steel.

Light is emitted from a source in the form of packets of waves. Light coming out from an ordinary source contains a mixture of many wavelengths. There is also no phase relation between the various waves. Therefore, such light, even if it is passed through an aperture, spreads very fast and the beam size increases rapidly with distance. In the case of laser light, the wavelength of each packet is almost the same. Also the average length of the packet of waves is much larger. This means that there is better phase correlation over a longer duration of time. This results in reducing the divergence of a laser beam substantially.

If there are N atoms in a source, each emitting light with intensity I, then the total intensity produced by an ordinary source is proportional to NI, whereas in a laser source, it is proportional to N^2I. Considering that N is very large, we see that the light from a laser can be much stronger than that from an ordinary source.

When astronauts of the Apollo missions visited the moon, they placed a mirror on its surface, facing the earth. Then scientists on the earth sent a strong laser beam, which was reflected by the mirror on the moon and received back on the earth. The size of the reflected laser beam and the time taken for the round trip were measured. This allowed a very accurate determination of (a) the extremely small divergence of a laser beam and (b) the distance of the moon from the earth.

SUMMARY

1. Atom, as a whole, is electrically neutral and therefore contains equal amount of positive and negative charges.
2. In *Thomson's model*, an atom is a spherical cloud of positive charges with electrons embedded in it.
3. In *Rutherford's model*, most of the mass of the atom and all its positive charge are concentrated in a tiny nucleus (typically one by ten thousand the size of an atom), and the electrons revolve around it.
4. Rutherford nuclear model has two main difficulties in explaining the structure of atom: (a) It predicts that atoms are unstable because the accelerated electrons revolving around the nucleus must spiral into the nucleus. This contradicts the stability of matter. (b) It cannot explain the characteristic line spectra of atoms of different elements.
5. Atoms of each element are stable and emit characteristic spectrum. The spectrum consists of a set of isolated parallel lines termed as line spectrum. It provides useful information about the atomic structure.
6. The atomic hydrogen emits a line spectrum consisting of various series. The frequency of any line in a series can be expressed as a difference of two terms;

Lyman series : $\nu = Rc\left(\frac{1}{1^2} - \frac{1}{n^2}\right)$; $n = 2, 3, 4, ...$

Balmer series : $\nu = Rc\left(\frac{1}{2^2} - \frac{1}{n^2}\right)$; $n = 3, 4, 5, ...$

Paschenseries : $\nu = Rc\left(\frac{1}{3^2} - \frac{1}{n^2}\right)$; $n = 4, 5, 6, ...$

Brackett series : $\nu = Rc\left(\frac{1}{4^2} - \frac{1}{n^2}\right)$; $n = 5, 6, 7, ...$

Pfundseries : $\nu = Rc\left(\frac{1}{5^2} - \frac{1}{n^2}\right)$; n = 6, 7, 8, ...

7. To explain the line spectra emitted by atoms, as well as the stability of atoms, Niel's Bohr proposed a model for hydrogenic (single elctron) atoms. He introduced three postulates and laid the foundations of quantum mechanics :
   (a) In a hydrogen atom, an electron revolves in certain stable orbits (called stationary orbits) without the emission of radiant energy.
   (b) The stationary orbits are those for which the angular momentum is some integral multiple of $h/2\pi$. (Bohr's quantisation condition.) That is $L = nh/2\pi$, where n is an integer called a quantum number.
   (c) The third postulate states that an electron might make a transition from one of its specified non-radiating orbits to another of lower energy. When it does so, a photon is emitted having energy equal to the energy difference between the initial and final states. The frequency (ν) of the emitted photon is then given by

   $$h\nu = E_i - E_f$$

   An atom absorbs radiation of the same frequency the atom emits, in which case the electron is transferred to an orbit with a higher value of n.

   $$E_i + h\nu = E_f$$

8. As a result of the quantisation condition of angular momentum, the electron orbits the nucleus at only specific radii. For a hydrogen atom it is given by

   $$r_n = \left(\frac{n^2}{m}\right)\left(\frac{h}{2\pi}\right)^2 \frac{4\pi\varepsilon_0}{e^2}$$

   The total energy is also quantised:

   $$E_n = \frac{-me^4}{8n^2\varepsilon_0^2 h^2}$$

   $$= -13.6 \text{ eV}/n^2$$

   The n = 1 state is called ground state. In hydrogen atom the ground state energy is –13.6 eV. Higher values of n correspond to excited states (n >1). Atoms are excited to these higher states by collisions with other atoms or electrons or by absorption of a photon of right frequency.

9. de Broglie's hypothesis that electrons have a wavelength $\lambda = h/mv$ gave an explanation for Bohr's quantised orbits by bringing in the wave-particle duality. The orbits correspond to circular standing waves in which the circumference of the orbit equals a whole number of wavelengths.

10. Bohr's model is applicable only to hydrogenic (single electron) atoms. It cannot be extended to even two electron atoms such as helium. This model is also unable to explain for the relative intensities of the frequencies emitted even by hydrogenic atoms.

VERY SHORT ANSWER QUESTIONS (2 Marks)

1. What is the angular momentum of electron in the second orbit of Bohr's model of hydrogen atom?
2. What is the expression for fine structure constant and what is its value?
3. What is the physical meaning of 'negative energy of an electron?
4. Sharp lines are present in the spectrum of a gas. What does this indicate?
5. Name a physical quantity whose dimensions are the same as those of angular momentum.
6. What is the difference between α-particle and helium atom?
7. Among alpha, beta and gamma radiation, which get affected by the electric field?
8. The Lyman series of hydrogen spectrum lies in the ultraviolet region. Why?
9. Give two drawbacks of Rutherford's atomic model.

SHORT ANSWER QUESTIONS (4 Marks)

1. Derive an expression for potential and kinetic energy of an electron in any orbit of a hydrogen atom according to Bohr's atomic model. How does, P.E. change with increasing n (orbit number)?
2. What are the limitations of Bohr's theory of hydrogen atom?
3. Explain the distance of closest approach and impact parameter.
4. Give a brief account of Thomson model of atom. What are its limitations?
5. Describe Rutherford atom model. What are the draw backs of this model?
6. Explain the different types of spectral series.
7. Write a short note on Debroglie's explanation of Bohr's second postulate of quantization.

LONG ANSWER QUESTIONS (8 Marks)

1. Describe Geiger-Marsden experiment on scattering of α-particles. How is the size of the nucleus estimated in this experiment?
2. Describe Bohr's theory of the spectrum of hydrogen atom.
3. State the basic postulates of Bohr's theory of atoms spectra. Hence obtain an expression for the radius of orbit and the energy of orbital electron is a hydrogen atom.

CHAPTER 14

NUCLEI

14.1 INTRODUCTION

In the previous chapter, we have learnt that in every atom, the positive charge and mass are densely concentrated at the centre of the atom forming its nucleus. The overall dimensions of a nucleus are much smaller than those of an atom. Experiments on scattering of α-particles demonstrated that the radius of a nucleus was smaller than the radius of an atom by a factor of about 10^4. This means the volume of a nucleus is about 10^{-12} times the volume of the atom. In other words, an atom is almost empty. If an atom is enlarged to the size of a classroom, the nucleus would be of the size of pinhead. Nevertheless, the nucleus contains most (more than 99.9%) of the mass of an atom.

Does the nucleus have a structure, just as the atom does? If so, what are the constituents of the nucleus? How are these held together? In this chapter, we shall look for answers to such questions. We shall discuss various properties of nuclei such as their size, mass and stability, and also associated nuclear phenomena such as radioactivity, fission and fusion.

14.2 ATOMIC MASSES AND COMPOSITION OF NUCLEUS

The mass of an atom is very small, compared to a kilogram; for example, the mass of a carbon atom, ^{12}C, is 1.992647×10^{-26} kg. Kilogram is not a very convenient unit to measure such small quantities. Therefore, a different mass unit is used for expressing atomic masses. This unit is the atomic mass unit (u), defined as $1/12^{th}$ of the mass of the carbon (^{12}C) atom. According to this definition

$$1\text{ u} = \frac{\text{mass of one } ^{12}\text{C atom}}{12}$$

$$= \frac{1.992648\times10^{-26}\text{ kg}}{12}$$

$$= 1.660539 \times 10^{-27}\text{ kg} \qquad (14.1)$$

The atomic masses of various elements expressed in atomic mass unit (u) are close to being integral multiples of the mass of a hydrogen atom. There are, however, many striking exceptions to this rule. For example, the atomic mass of chlorine atom is 35.46 u.

Accurate measurement of atomic masses is carried out with a mass spectrometer, the measurement of atomic masses reveals the existence of different types of atoms of the same element, which exhibit the same chemical properties, but differ in mass. Such atomic species of the same element differing in mass are called *isotopes*. (In Greek, isotope means the same place, i.e. they occur in the same place in the periodic table of elements.) It was found that practically every element consists of a mixture of several isotopes. The relative abundance of different isotopes differs from element to element. Chlorine, for example, has two isotopes having masses 34.98 u and 36.98 u, which are nearly integral multiples of the mass of a hydrogen atom. The relative abundances of these isotopes are 75.4 and 24.6 per cent, respectively. Thus, the average mass of a chlorine atom is obtained by the weighted average of the masses of the two isotopes, which works out to be

$$= \frac{75.4\times34.98 + 24.6\times36.98}{100}$$

$$= 35.47\text{ u}$$

which agrees with the atomic mass of chlorine.

Even the lightest element, hydrogen has three isotopes having masses 1.0078 u, 2.0141 u, and 3.0160 u. The nucleus of the lightest atom of hydrogen, which has a relative abundance of 99.985%, is called the proton. The mass of a proton is

$$m_p = 1.00727 \text{ u} = 1.67262 \times 10^{-27} \text{ kg} \qquad (14.2)$$

This is equal to the mass of the hydrogen atom (= 1.00783u), minus the mass of a single electron (m_e= 0.00055 u). The other two isotopes of hydrogen are called deuterium and tritium. Tritium nuclei, being unstable, do not occur naturally and are produced artificially in laboratories.

The positive charge in the nucleus is that of the protons. A proton carries one unit of fundamental charge and is stable. It was earlier thought that the nucleus may contain electrons, but this was ruled out later using arguments based on quantum theory. All the electrons of an atom are outside the nucleus. We know that the number of these electrons outside the nucleus of the atom is *Z*, the atomic number. The total charge of the atomic electrons is thus (–*Ze*), and since the atom is neutral, the charge of the nucleus is (+*Ze*). The number of protons in the nucleus of the atom is, therefore, exactly *Z*, the atomic number.

Discovery of Neutron

Since the nuclei of deuterium and tritium are isotopes of hydrogen, they must contain only one proton each. But the masses of the nuclei of hydrogen, deuterium and tritium are in the ratio of 1:2:3. Therefore, the nuclei of deuterium and tritium must contain, in addition to a proton, some neutral matter. The amount of neutral matter present in the nuclei of these isotopes, expressed in units of mass of a proton, is approximately equal to one and two, respectively. This fact indicates that the nuclei of atoms contain, in addition to protons, neutral matter in multiples of a basic unit. This hypothesis was verified in 1932 by James Chadwick who observed emission of neutral radiation when beryllium nuclei were bombarded with alpha-particles (α-particles are helium nuclei, to be discussed in a later section). It was found that this neutral radiation could knock out protons from light nuclei such as those of helium, carbon and nitrogen. The only neutral radiation known at that time was photons (electromagnetic radiation). Application of the principles of conservation of energy and momentum showed that if the neutral radiation consisted of photons, the energy of photons would have to be much higher than is available from the bombardment of beryllium nuclei with α-particles. The clue to this puzzle, which Chadwick satisfactorily solved, was to assume that the neutral radiation consists of a new type of neutral particles called *neutrons*. From conservation of energy and momentum, he was able to determine the mass of new particle 'as very nearly the same as mass of proton'.

The mass of a neutron is now known to a high degree of accuracy. It is

$$m_n = 1.00866 \text{ u} = 1.6749 \times 10^{-27} \text{ kg} \qquad (14.3)$$

Chadwick was awarded the 1935 Nobel Prize in Physics for his discovery of the neutron.

A free neutron, unlike a free proton, is unstable. It decays into a proton, an electron and a antineutrino (another elementary particle), and has a mean life of about 1000s. It is, however, stable inside the nucleus.

The composition of a nucleus can now be described using the following terms and symbols :

Z -atomic number = number of protons [14.4(a)]

N - neutron number = number of neutrons [14.4(b)]

A - mass number = *Z* + *N* = total number of protons and neutrons [14.4(c)]

One also uses the term nucleon for a proton or a neutron. Thus the number of nucleons in an atom is its mass number A.

Nuclear species or nuclides are shown by the notation ${}_{Z}X^{A}$ where X is the chemical symbol of the species. For example, the nucleus of gold is denoted by ${}_{79}Au^{197}$. It contains 197 nucleons, of which 79 are protons and the rest118 are neutrons.

The composition of isotopes of an element can now be readily explained. The nuclei of isotopes of a given element contain the same number of protons, but differ from each other in their number of neutrons. Deuterium, ${}_{1}H^{2}$, which is an isotope of hydrogen, contains one proton and one neutron. Its other isotope tritium, ${}_{1}H^{2}$, contains one proton and two neutrons. The element gold has 32 isotopes, ranging from A =173 to A = 204. We have already mentioned that chemical properties of elements depend on their electronic structure. As the atoms of isotopes have identical electronic structure they have identical chemical behaviour and are placed in the same location in the periodic table.

All nuclides with same mass number *A*are called *isobars*. For example, the nuclides ${}_{1}H^{3}$ and ${}_{2}He^{3}$ are isobars. Nuclides with same neutron number *N* but different atomic number *Z*, for example ${}_{90}Hg^{198}$ and ${}_{79}Au^{197}$, are called *isotones*.

14.3 SIZE OF THE NUCLEUS

As we have seen in Chapter 13, Rutherford was the pioneer who postulated and established the existence of the atomic nucleus. At Rutherford's suggestion, Geiger and Marsden performed their classic experiment: on the scattering of α-particles from thin gold foils. Their experiments revealed that the distance of closest approach to a gold nucleus of anα-particle of kinetic energy 5.5 MeV is about 4.0×10^{-14}m. The scattering of α-particle by the gold sheet could be understood by Rutherford by assuming that the coulomb repulsive force was solely responsible for scattering. Since the positive charge is confined to the nucleus, the actual size of the nucleus has to be less than 4.0×10^{-14}m. If we use α-particles of higher energies than 5.5 MeV, the distance of closest approach to the gold nucleus will be smaller and at some point the scattering will begin to be affected by the short range nuclear forces, and differ from Rutherford's calculations. Rutherford's calculations are based on pure coulomb repulsion between the positive charges of the α- particle and the gold nucleus. From the distance at which deviations set in, nuclear sizes can be inferred.

By performing scattering experiments in which fast electrons, instead of α-particles, are projectiles that bombard targets made up of various elements, the sizes of nuclei of various elements have been accurately measured.

It has been found that a nucleus of mass number *A* has a radius

$$R = R_0 \, A^{1/3} \qquad (14.5)$$

whereR_0 = 1.2×10^{-15} m (=1.2 fm; 1 fm = 10^{-15} m). This means the volume of the nucleus, which is proportional to R^3 is proportional to *A*. Thus the density of nucleus is a constant, independent of *A,* for all nuclei. Different nuclei are like a drop of liquid of constant density. The density of nuclear matter is approximately 2.3×10^{17} kg m^{-3}. This density is very large compared to ordinary matter, say water, which is 10^3 kg m^{-3}. This is understandable, as we have already seen that most of the atom is empty. Ordinary matter consisting of atoms has a large amount of empty space.

14.4 MASS-ENERGY AND NUCLEAR BINDING ENERGY

14.4.1 Mass – Energy

Einstein showed from his theory of special relativity that it is necessary to treat mass as another form of energy. Before the advent of this theory of special relativity it was presumed that mass and energy were conserved separately in a reaction. However, Einstein showed that mass is another form of energy and one can convert mass-energy into other forms of energy, say kinetic energy and vice-versa.

Einstein gave the famous mass-energy equivalence relation

$$E = mc^2 \qquad (14.6)$$

Here the energy equivalent of mass m is related by the above equation and c is the velocity of light in vacuum and is approximately equal to 3×10^8 m s^{-1}.

Experimental verification of the Einstein's mass-energy relation has been achieved in the study of nuclear reactions amongst nucleons, nuclei, electrons and other more recently discovered particles. In a reaction the conservation law of energy states that the initial energy and the final energy are equal provided the energy associated with mass is also included. This concept is important in understanding nuclear masses and the interaction of nuclei with one another. They form the subject matter of the next few sections.

14.4.2 Nuclear binding energy

In Section 14.2 we have seen that the nucleus is made up of neutrons and protons. Therefore it may be expected that the mass of the nucleus is equal to the total mass of its individual protons and neutrons. However, the nuclear mass M is found to be always less than this. For example, let us consider ${}_8O^{16}$; a nucleus which has 8 neutrons and 8 protons. We have

Mass of 8 neutrons = 8 × 1.00866 u

Mass of 8 protons = 8 × 1.00727 u

Mass of 8 electrons = 8 × 0.00055 u

Therefore the expected mass of ${}_8O^{16}$ nucleus

= 8 × 2.01593 u = 16.12744 u.

The atomic mass of ${}_8O^{16}$ found from mass spectroscopy experiments is seen to be 15.99493 u. Subtracting the mass of 8 electrons (8 × 0.00055 u) from this, we get the experimental mass of ${}_8O^{16}$ nucleus to be 15.99053 u.

Thus, we find that the mass of the ${}_8O^{16}$ nucleus is less than the total mass of its constituents by 0.13691u. The difference in mass of a nucleus and its constituents, ΔM, is called the *mass defect*, and is given by

$$\Delta M = [Zm_p + (A - Z)m_n] - M \qquad (14.7)$$

What is the meaning of the mass defect? It is here that Einstein's equivalence of mass and energy plays a role. Since the mass of the oxygen nucleus is less that the sum of the masses of its constituents (8 protons and 8 neutrons, in the unbound state), the equivalent energy of the oxygen nucleus is less than that of the sum of the equivalent energies of its constituents. If one wants to break the oxygen nucleus into 8 protons and 8 neutrons, this extra energy ΔMc^2, has to supplied. This energy required Eb is related to the mass defect by

$$E_b = \Delta M\, c^2 \qquad (14.8)$$

If a certain number of neutrons and protons are brought together to form a nucleus of a certain charge and mass, an energy E_b will be released in the process. The energy E_b is called the *binding energy* of the nucleus. If we separate a nucleus into its nucleons, we would have

to supply a total energy equal to E_b, to those particles. Although we cannot tear apart a nucleus in this way, the nuclear binding energy is still a convenient measure of how well a nucleus is held together. A more useful measure of the binding between the constituents of the nucleus is the *binding energy per nucleon*, E_{bn}, which is the ratio of the binding energy E_b of a nucleus to the number of the nucleons, A, in that nucleus :

$$E_{bn} = E_b / A \qquad (14.9)$$

We can think of binding energy per nucleon as the average energy per nucleon needed to separate a nucleus into its individual nucleons.

Figure 14.1 is a plot of the binding energy per nucleon E_{bn} versus the mass number A for a large number of nuclei. We notice the following main features of the plot :

(i) the binding energyper nucleon, E_{bn}, is practically constant, i.e. practically independent of the atomic number for nuclei of middle mass number (30 < A < 170). The curve has a maximum of about 8.75 MeV for A = 56 and has a value of 7.6 MeV for A = 238.

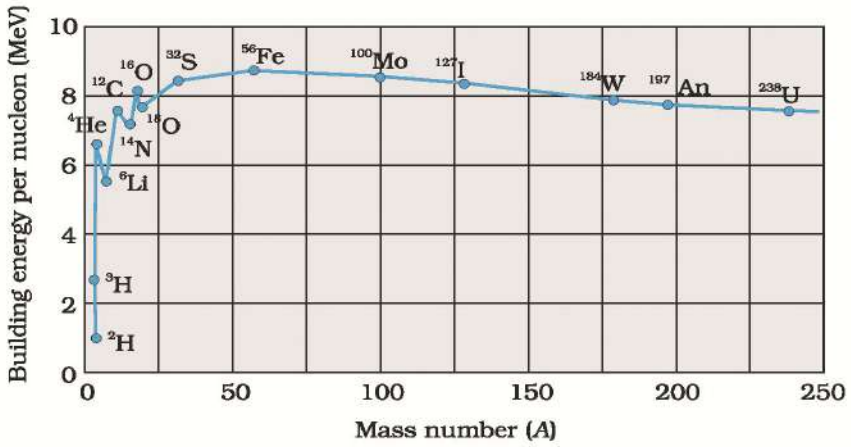


Fig. 14.1 The binding energy per nucleon as a function of mass number.

(ii) E_{bn} is lower for both light nuclei (A<30) and heavy nuclei (A>170).

We can draw some conclusions from these two observations:

(i) The force is attractive and sufficiently strong to produce a binding energy of a few MeV per nucleon.

(ii) The constancy of the binding energy in the range 30 < A < 170 is a consequence of the fact that the nuclear force is short-ranged. Consider a particular nucleon inside a sufficiently large nucleus. It will be under the influence of only some of its neighbours, which come within the range of the nuclear force. If any other nucleon is at a distance more than the range of the nuclear force from the particular nucleon it will have no influence on the binding energy of the nucleon under consideration. If a nucleon can have a maximum of p neighbours within the range of nuclear force, its binding energy would be proportional to p. Let the binding energy of the nucleus be pk, where k is a constant having the dimensions of energy. If we increase A by adding nucleons they will not change the binding energy of a nucleon inside. Since most of the nucleons in a large nucleus reside inside it and not on the surface, the change in binding energy per nucleon would be small. The binding energy per nucleon is a constant and is approximately equal to pk. The property that a given nucleon influences only nucleons close to it is also referred to as saturation property of the nuclear force.

(iii) A very heavy nucleus, say A = 240, has lower binding energy per nucleon compared to that of a nucleus with A = 120. Thus if a nucleus A = 240 breaks into two A = 120

nuclei, nucleons get more tightly bound. This implies energy would be released in the process. It has very important implications for energy production through *fission*, to be discussed later in Section 14.7.1.

(iv) Consider two very light nuclei ($A \leq 10$) joining to form a heavier nucleus. The binding energy per nucleon of the fused heavier nuclei is more than the binding energy per nucleon of the lighter nuclei. This means that the final system is more tightly bound than the initial system. Again energy would be released in such a process of *fusion*. This is the energy source of sun, to be discussed later in Section 14.7.3.

14.5 NUCLEAR FORCE

The force that determines the motion of atomic electrons is the familiar Coulomb force. In Section 14.4, we have seen that for average mass nuclei the binding energy per nucleon is approximately 8 MeV, which is much larger than the binding energy in atoms. Therefore, to bind a nucleus together there must be a strong attractive force of a totally different kind. It must be strong enough to overcome the repulsion between the (positively charged) protons and to bind both protons and neutrons into the tiny nuclear volume. We have already seen that the constancy of binding energy per nucleon can be understood in terms of its short-range. Many features of the nuclear binding force are summarised below. These are obtained from a variety of experiments carried out during 1930 to 1950.

(i) The nuclear force is much stronger than the Coulomb force acting between charges or the gravitational forces between masses. The nuclear binding force has to dominate over the Coulomb repulsive force between protons inside the nucleus. This happens only because the nuclear force is much stronger than the coulomb force. The gravitational force is much weaker than even Coulomb force.

(ii) The nuclear force between two nucleons falls rapidly to zero as their distance is more than a few femtometres. This leads to *saturation of forces* in a medium or a large-sized nucleus, which is the reason for the constancy of the binding energy per nucleon.

A rough plot of the potential energy between two nucleons as a function of distance is shown in the Fig. 14.2. The potential energy is a minimum at a distance r_0 of about 0.8 fm. This means that the force is attractive for distances larger than 0.8 fm and repulsive if they are separated by distances less than 0.8 fm.

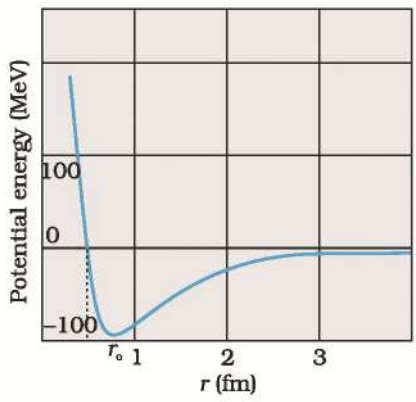


Fig. 14.2 Potential energy of a pair of nucleons as a function of their separation. For a separation greater than r_0, the force is attractive and for separations less than r_0, the force is strongly repulsive.

(iii) The nuclear force between neutron-neutron, proton-neutron and proton-proton is approximately the same. The nuclear force does not depend on the electric charge.

Unlike Coulomb's law or the Newton's law of gravitation there is no simple mathematical form of the nuclear force.

14.6 RADIOACTIVITY

A.H. Becquerel discovered radioactivity in 1896 purely by accident. While studying the fluorescence and phosphorescence of compounds irradiated with visible light, Becquerel observed an interesting phenomenon. After illuminating some pieces of uranium-potassium sulphate with visible light, he wrapped them in black paper and separated the package from a photographic plate by a piece of silver. When, after several hours of exposure, the photographic plate was developed, it showed

blackening due to something that must have been emitted by the compound and was able to penetrate both black paper and the silver.

Experiments performed subsequently showed that radioactivity was a nuclear phenomenon in which an unstable nucleus undergoes a decay. This is referred to as *radioactive decay*. Three types of radioactive decay occur in nature:

(i) α-decay in which a helium nucleus 4 He is emitted;

(ii) β-decay in which electrons or positrons (particles with the same mass as electrons, but with a charge exactly opposite to that of electron) are emitted;

(iii) γ-decay in which high energy (hundreds of keV or more) photons are emitted.

Each of these decay will be considered in subsequent sub-sections.

14.6.1 Law of radioactive decay

In any radioactive sample, which undergoes α, β or γ-decay, it is found that the number of nuclei undergoing the decay per unit time is proportional to the total number of nuclei in the sample. If N is the number of nuclei in the sample and ΔN undergo decay in time Δt then

$$\frac{\Delta N}{\Delta t} \propto N$$

or, $\Delta N/\Delta t = \lambda N$, (14.10)

where λ is called the radioactive *decay constant* or *disintegration constant*.

The change in the number of nuclei in the sample* is $dN = -\Delta N$ in time Δt. Thus the rate of change of N is (in the limit $\Delta t \to 0$)

$$\frac{dN}{dt} = -\lambda N$$

* ΔN is the number of nuclei that decay, and hence is always positive. dN is the change in N, which may have either sign. Here it is negative, because out of original N nuclei, ΔN have decayed, leaving $(N - \Delta N)$ nuclei.

or, $$\frac{dN}{N} = -\lambda dt$$

Now, integrating both sides of the above equation, we get,

$$\int_{N_0}^{N} \frac{dN}{N} = -\lambda \int_{t_0}^{t} dt \qquad (14.11)$$

or, $$\ln N - \ln N_0 = -\lambda(t - t_0) \qquad (14.12)$$

Here N_0 is the number of radioactive nuclei in the sample at some arbitrary time t_0 and N is the number of radioactive nuclei at any subsequent time t. Setting $t_0 = 0$ and rearranging Eq. (14.12) gives us

$$\ln \frac{N}{N_0} = -\lambda t \qquad (14.13)$$

which gives

$$N(t) = N_0 \, e^{-\lambda t} \qquad (14.14)$$

Note, for example, the light bulbs follow no such exponential decay law. If we test 1000 bulbs for their life (time span before they burn out or fuse), we expect that they will 'decay' (that is, burn out) at more or less the same time. The decay of radionuclides follows quite a different law, the *law of radioactive decay* represented by Eq. (14.14).

The total decay rate R of a sample is the number of nuclei disintegrating per unit time. Suppose in a time interval dt, the decay count measured is ΔN. Then $dN = -\Delta N$.

The positive quantity R is then defined as

$$R = -\frac{dN}{dt}$$

Differentiating Eq. (14.14), we get

$$R = \lambda N e^{-\lambda t}$$

or, $$R = R_0 e^{-\lambda t} \quad (14.15)$$

This is equivalent to the law of radioactivity decay, since you can integrate Eq. (14.15) to get back Eq. (14.14). Clearly, $R_0 = \lambda N_0$ is the decay rate at $t = 0$. The decay rate R at a certain time t and the number of undecayed nuclei N at the same time are related by

$$R = \lambda N \quad (14.16)$$

The decay rate of a sample, rather than the number of radioactive nuclei, is a more direct experimentally measurable quantity and is given a specific name: *activity*. The SI unit for activity is Becquerel, named after the discoverer of radioactivity, Henry Becquerel.

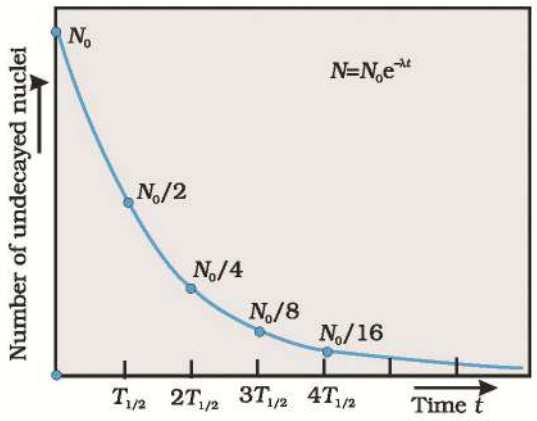


Fig. 13.3 Exponential decay of a radioactive species. After a lapse of $T_{1/2}$, population of the given species drops by a factor of 2.

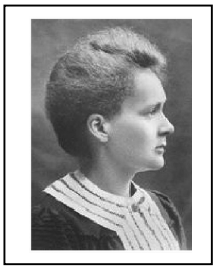
1 becquerel is simply equal to 1 disintegration or decay per second. There is also another unit named "curie" that is widely used and is related to the SI unit as:

1 curie = 1 Ci = 3.7×10^{10} decays per second

$= 3.7 \times 10^{10}$ Bq

Different radionuclides differ greatly in their rate of decay. A common way to characterize this feature is through the notion of *half-life*. Half-life of a radionuclide (denoted by $T_{1/2}$) is the time it takes for a sample that has initially, say N_0 radionuclei to reduce to $N_0/2$. Putting $N = N_0/2$ and $t = T_{1/2}$ in Eq. (14.14), we get

$$T_1/2 = \frac{\ln 2}{\lambda} = \frac{0.693}{\lambda} \quad (14.17)$$



MARIE SKLODOWSKA CURIE (1867-1934)

Marie Sklodowska Curie (1867-1934) Born in Poland. She is recognised both as a physicist and as a chemist. The discovery of radioactivity by Henri Becquerel in 1896 inspired Marie and her husband Pierre Curie in their researches and analyses which led to the isolation of radium and polonium elements. She was the first person to be awarded two Nobel Prizes- for Physics in 1903 and for Chemistry in 1911.

Clearly if N_0 reduces to half its value in time $T_{1/2}$, R_0 will also reduce to half its value in the same time according to Eq. (14.16).

Another related measure is the *average* or *mean life* τ. This again can be obtained from Eq. (14.14). The number of nuclei which decay in the time interval t to $t + \Delta t$ is $R(t)\Delta t$ ($= \lambda N e^{-\lambda t}\Delta t$). Each of them has lived for time t. Thus the total life of all these nuclei would be t

$\lambda N_0 e^{-\lambda t}\Delta t$. It is clear that some nuclei may live for a short time while others may live longer. Therefore to obtain the mean life, we have to sum (or integrate) this expression over all times from 0 to ∞, and divide by the total number $N0$ of nuclei at $t = 0$. Thus,

$$\tau = \frac{\lambda N_0 \int_0^\infty t e^{-\lambda t} dt}{N_0} = \lambda \int_0^\infty t e^{-\lambda t} dt$$

One can show by performing this integral that

$\tau = 1/\lambda$

We summarise these results with the following:

$$T_{1/2} = \frac{\ln 2}{\lambda} = \tau \ln 2 \qquad (14.18)$$

Radioactive elements (e.g., tritium, plutonium) which are short-lived i.e., have half-lives much less than the age of the universe (–15 billion years) have obviously decayed long ago and are not found in nature. They can, however, be produced artificially in nuclear reactions.

14.6.2 Alpha decay

A well-known example of alpha decay is the decay of uranium ${}_{92}U^{238}$ to thorium ${}_{90}Th^{234}$ with the emission of a helium nucleus ${}_{2}He^{4}$

$${}_{90}U^{238} \rightarrow {}_{90}Th^{234} + {}_{2}He^{4} \qquad (\alpha\text{-decay}) \qquad (14.19)$$

In α-decay, the mass number of the product nucleus (daughter nucleus) is four less than that of the decaying nucleus (parent nucleus), while the atomic number decreases by two. In general, α-decay of a parent nucleus ${}_{Z}X^{A}$ results in a daughter nucleus ${}_{Z-2}Y^{A-4}$

$${}_{Z}X^{A} \rightarrow {}_{Z-2}Y^{A-4} + {}_{2}He^{4} \qquad (14.20)$$

From Einstein's mass-energy equivalance relation [Eq. (14.6)] and energy conservation, it is clear that this spontaneous decay is possible only when the total mass of the decay products is less than the mass of the initial nucleus. This difference in mass appears as kinetic energy of the products. By referring to a table of nuclear masses, one can check that the total mass of ${}_{90}Th^{234}$ and ${}_{2}He^{4}$ is indeed less than that of ${}_{92}U^{238}$.

The disintegration energy or the Q-value of a nuclear reaction is the difference between the initial mass energy and the total mass energy of the decay products. For α-decay

$$Q = (m_x - m_y - m_{He})\, c^2 \qquad (14.21)$$

Q is also the net kinetic energy gained in the process or, if the initial nucleus X is at rest, the kinetic energy of the products. Clearly, $Q > 0$ for exothermic processes such as α-decay.

14.6.3 Beta decay

In beta decay, a nucleus spontaneously emits an electron (β^- decay) or a positron (β^+ decay). A common example of β^- decay is

$${}_{15}P^{32} \rightarrow {}_{16}S^{32} + e^- + \bar{\nu} \qquad (14.22)$$

and that of β^+ decay is

$${}_{11}Na^{22} \rightarrow {}_{10}Ne^{22} + e^+ + \nu \qquad (14.23)$$

The decays are governed by the Eqs. (14.14) and (14.15), so that one can never predict which *nucleus* will undergo decay, but one can characterize the decay by a half-life $T_{1/2}$. For

example, $T_{1/2}$ for the decays above is respectively 14.3 d and 2.6y. The emission of electron in β^- decay is accompanied by the emission of an antineutrino ($\bar{\nu}$); in β^+decay, instead, a neutrino (ν) is generated. Neutrinos are neutral particles with very small (possiblly, even zero) mass compared to electrons. They have only weak interaction with other particles. They are, therefore, very difficult to detect, since they can penetrate large quantity of matter (even earth) without any interaction.

In both β^- and β^+ decay, the mass number A remains unchanged. In β^- decay, the atomic number Z of the nucleus goes up by 1, while in β^+ decay Z goes down by 1. The basic nuclear process underlying β^- decay is the conversion of neutron to proton

$$n \rightarrow p + e^- + \bar{\nu} \qquad (14.24)$$

while for β^+ decay, it is the conversion of proton into neutron

$$p \rightarrow n + e^+ + \nu \qquad (14.25)$$

Note that while a free neutron decays to proton, the decay of proton to neutron [Eq. (14.25)] is possible only inside the nucleus, since proton has smaller mass than neutron.

14.6.4 Gamma decay

Like an atom, a nucleus also has discrete energy levels - the ground state and excited states. The scale of energy is, however, very different. Atomic energy level spacings are of the order of eV, while the difference in nuclear energy levels is of the order of MeV. When a nucleus in an excited state spontaneously decays to its ground state (or to a lower energy state), a photon is emitted with energy equal to the difference in the two energy levels of the nucleus. This is the so-called *gamma decay*. The energy (MeV) corresponds to radiation of extremely short wavelength, shorter than the hard X-ray region.

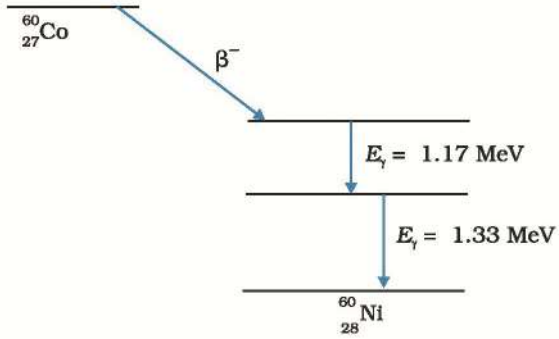


Fig. 14.4 β-decay of $_{28}Ni^{60}$ nucleus followed by emission of two γrays from de-excitation of the daughter nucleus$_{28}Ni^{90}$.

Typically, a gamma ray is emitted when a α or β decay results in a daughter nucleus in an excited state. This then returns to the ground state by a single photon transition or successive transitions involving more than one photon. A familiar example is the successive emission of gamma rays of energies 1.17 MeV and 1.33 MeV from the de-excitation of $_{28}Ni^{60}$ nuclei formed from β^- decay of $_{27}Co^{60}$.

14.7 NUCLEAR ENERGY

The curve of binding energy per nucleon E_{bn}, given in Fig. 14.1, has a long flat middle region between A = 30 and A = 170. In this region the binding energy per nucleon is nearly constant (8.0 MeV). For the lighter nuclei region, A < 30, and for the heavier nuclei region, A > 170, the binding energy per nucleon is less than 8.0 MeV, as we have noted earlier. Now, the greater the binding energy, the less is the total mass of a bound system, such as a nucleus. Consequently, if nuclei with less total binding energy transform to nuclei with greater binding energy, there will be a net energy release. This is what happens when a heavy nucleus decays into two or more intermediate mass fragments (*fission*) or when light nuclei fuse into a heavier nucleus (*fusion*.)

Exothermic chemical reactions underlie conventional energy sources such as coal or petroleum. Here the energies involved are in the range of electron volts. On the other hand, in a nuclear reaction, the energy release is of the order of MeV. Thus for the same quantity of matter, nuclear sources produce a million times more energy than a chemical source. Fission of 1 kg of uranium, for example, generates 10^{14} J of energy; compare it with burning of 1 kg of coal that gives 10^7 J.

14.7.1 Fission

New possibilities emerge when we go beyond natural radioactive decays and study nuclear reactions by bombarding nuclei with other nuclear particles such as proton, neutron, α-particle, etc.

A most important neutron-induced nuclear reaction is fission. An example of fission is when a uranium isotope ${}_{92}U^{235}$ bombarded with a neutron breaks into two intermediate mass nuclear fragments

$$ {}_0n^1 + {}_{92}U^{235} \rightarrow {}_{92}U^{236} \rightarrow {}_{56}Ba^{144} + {}_{36}Kr^{89} + 3\ {}_0n^1 \qquad (14.26) $$

The same reaction can produce other pairs of intermediate mass fragments

$$ {}_0n^1 + {}_{92}U^{235} \rightarrow {}_{92}U^{236} \rightarrow {}_{51}Sb^{133} + {}_{41}Nb^{99} + 4\ {}_0n^1 \qquad (14.27) $$

Or, as another example,

$$ {}_0n^1 + {}_{92}U^{235} \rightarrow {}_{54}Xe^{140} + {}_{38}Sr^{94} + 2\ {}_0n^1 \qquad (14.28) $$

The fragment products are radioactive nuclei; they emit βparticles in succession to achieve stable end products.

The energy released (the Q value) in the fission reaction of nuclei like uranium is of the order of 200 MeV per fissioning nucleus. This is estimated as follows:

Let us take a nucleus with A = 240 breaking into two fragments each of A = 120. Then

$E_{bn}$for A = 240 nucleus is about 7.6 MeV,

$E_{bn}$for the two A = 120 fragment nuclei is about 8.5 MeV.

$\therefore$ Gain in binding energy for nucleon is about 0.9 MeV. Hence the total gain in binding energy is 240×0.9 or 216 MeV.

The disintegration energy in fission events first appears as the kinetic energy of the fragments and neutrons. Eventually it is transferred to the surrounding matter appearing as heat. The source of energy in nuclear reactors, which produce electricity, is nuclear fission. The enormous energy released in an atom bomb comes from uncontrolled nuclear fission. We discuss some details in the next section how a nuclear reactor functions.

INDIA'S ATOMIC ENERGY PROGRAMME

The atomic energy programme in India was launched around the time of independence under the leadership of Homi J. Bhabha (1909-1966). An early historic achievement was the design and construction of the first nuclear reactor in India (named Apsara) which went critical on August 4, 1956. It used enriched uranium as fuel and water as moderator. Following this was another notable landmark: the construction of CIRUS (Canada India Research U.S.) reactor in 1960. This 40 MW reactor used natural uranium as fuel and heavy water as moderator. Apsara and CIRUS spurred research in a wide range of areas of basic and applied nuclear science. An important milestone in the first two decades of the programme was the indigenous design and construction of the plutonium plant at Trombay, which ushered in the technology of fuel reprocessing (separating useful fissile and fertile nuclear materials from the spent fuel of a reactor) in India. Research reactors that have been subsequently commissioned include ZERLINA, PURNIMA (I, II and III), DHRUVA and

KAMINI. KAMINI is the country's first large research reactor that uses U-233 as fuel. As the name suggests, the primary objective of a research reactor is not generation of power but to provide a facility for research on different aspects of nuclear science and technology. Research reactors are also an excellent source for production of a variety of radioactive isotopes that find application in diverse fields: industry, medicine and agriculture.

The main objectives of the Indian Atomic Energy programme are to provide safe and reliable electric power for the country's social and economic progress and to be self- reliant in all aspects of nuclear technology. Exploration of atomic minerals in India undertaken since the early fifties has indicated that India has limited reserves of uranium, but fairly abundant reserves of thorium. Accordingly, our country has adopted a three- stage strategy of nuclear power generation. The first stage involves the use of natural uranium as a fuel, with heavy water as moderator. The Plutonium-239 obtained from reprocessing of the discharged fuel from the reactors then serves as a fuel for the second stage — the fast breeder reactors. They are so called because they use fast neutrons for sustaining the chain reaction (hence no moderator is needed) and, besides generating power, also breed more fissile species (plutonium) than they consume. The third stage, most significant in the long term, involves using fast breeder reactors to produce fissile Uranium-233 from Thorium-232 and to build power reactors based on them.

India is currently well into the second stage of the programme and considerable work has also been done on the third — the thorium utilisation — stage. The country has mastered the complex technologies of mineral exploration and mining, fuel fabrication, heavy water production, reactor design, construction and operation, fuel reprocessing, etc. Pressurised Heavy Water Reactors (PHWRs) built at different sites in the country mark the accomplishment of the first stage of the programme. India is now more than self-sufficient in heavy water production. Elaborate safety measures both in the design and operation of reactors, as also adhering to stringent standards of radiological protection are the hallmark of the Indian Atomic Energy Programme.

14.7.2 Nuclear reactor

Notice one fact of great importance in the fission reactions given in Eqs. (14.26) to (14.28). There is a release of *extra* neutron (s) in the fission process. Averagely, 2½ neutrons are released per fission of uranium nucleus. It is a fraction since in some fission events 2 neutrons are produced, in some 3, etc. The extra neutrons in turn can initiate fission processes, producing still more neutrons, and so on. This leads to the possibility of a chain reaction, as was first suggested by Enrico Fermi. If the chain reaction is controlled suitably, we can get a steady energy output. This is what happens in a nuclear reactor. If the chain reaction is uncontrolled, it leads to explosive energy output, as in a nuclear bomb. There is, however, a hurdle in sustaining a chain reaction, as described here. It is known experimentally that slow neutrons (thermal neutrons) are much more likely to cause fission in 235 U than fast neutrons. Also fast neutrons liberated in fission would escape instead of causing another fission reaction.

The average energy of a neutron produced in fission of $_{92}U^{235}$ is 2 MeV. These neutrons unless slowed down will escape from the reactor without interacting with the uranium nuclei, unless a very large amount of fissionable material is used for sustaining the chain reaction. What one needs to do is to slow down the fast neutrons by elastic scattering with light nuclei. In fact, Chadwick's experiments showed that in an elastic collision with hydrogen the neutron almost comes to rest and proton carries away the energy. This is the same situation as when a marble hits head-on an identical marble at rest. Therefore, in reactors, light nuclei called *moderators* are provided along with the fissionable nuclei for slowing down fast neutrons. The moderators commonly used are water, heavy water (D_2O) and graphite. The Apsara

reactor at the Bhabha Atomic Research Centre (BARC), Mumbai, uses water as moderator. The other Indian reactors, which are used for power production, use heavy water as moderator.

Because of the use of moderator, it is possible that the ratio, *K*, of number of fission produced by a given generation of neutrons to the number of fission of the preceding generation may be greater than one. This ratio is called the *multiplication factor;* it is the measure of the growth rate of the neutrons in the reactor. For $K = 1$, the operation of the reactor is said to be *critical*, which is what we wish it to be for steady power operation. If *K* becomes greater than one, the reaction rate and the reactor power increases exponentially. Unless the factor *K* is brought down very close to unity, the reactor will become supercritical and can even explode. The explosion of the Chernobyl reactor in Ukraine in 1986 is a sad reminder that accidents in a nuclear reactor can be catastrophic.

The reaction rate is controlled through control-rods made out of neutron-absorbing material such as cadmium. In addition to control rods, reactors are provided with *safety rods* which, when required, can be inserted into the reactor and *K* can be reduced rapidly to less than unity. The more abundant isotope ${}_{92}U^{238}$ in naturally occurring uranium is non-fissionable. When it captures a neutron, it produces the highly radioactive plutonium through these reactions

$$ {}_{92}U^{238} + n \rightarrow {}_{92}U^{239} \rightarrow {}_{93}Np^{239} + e^{-} +$$

$$ {}_{93}Np^{239} \rightarrow {}_{94}Pu^{239} + e^{-} + \qquad (14.29)$$

Plutonium undergoes fission with slow neutrons.

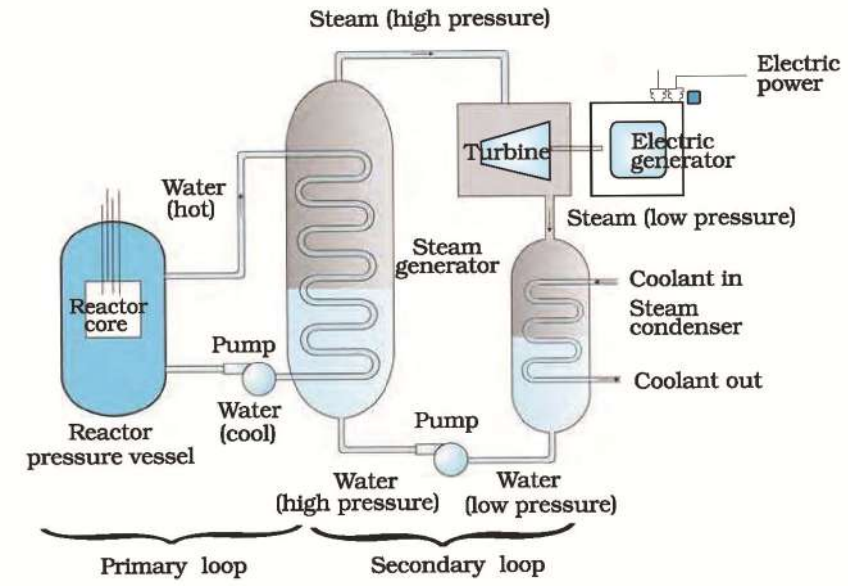


Fig. 14.5 Schematic diagram of a nuclear reactor based on thermal neutron fission.

Figure 14.5 shows the schematic diagram of a nuclear reactor based on thermal neutron fission. The *core* of the reactor is the site of nuclear fission. It contains the fuel elements in suitably fabricated form. The fuel may be say enriched uranium (i.e., one that has greater abundance of ${}_{92}U^{235}$ than naturally occurring uranium). The core contains a moderator to slow down the neutrons. The core is surrounded by a *reflector* to reduce leakage. The energy (heat) released in fission is continuously removed by a suitable *coolant*. A containment vessel prevents the escape of radioactive fission products. The whole assembly is shielded to check harmful radiation from coming out. The reactor can be shut down by means of rods (made of, for example, cadmium) that have high absorption of neutrons. The coolant transfers heat to a

working fluid which in turn may produce stream. The steam drives turbines and generates electricity. Like any power reactor, nuclear reactors generate considerable waste products. But nuclear wastes need special care for treatment since they are radioactive and hazardous. Elaborate safety measures, both for reactor operation as well as handling and reprocessing the spent fuel, are required. These safety measures are a distinguishing feature of the Indian Atomic Energy programme. An appropriate plan is being evolved to study the possibility of converting radioactive waste into less active and short-lived material.

14.7.3 Nuclear fusion – energy generation in stars

When two light nuclei fuse to form a larger nucleus, energy is released, since the larger nucleus is more tightly bound, as seen from the binding energy curve in Fig.14.1. Some examples of such energy liberating nuclear fusion reactions are :

$$_1H^1 + _1H^1 \rightarrow {}_1H^2 + e^+ + \nu + 0.42 \text{ MeV} \qquad [14.29(a)]$$

$$_1H^2 + _1H^2 \rightarrow {}_2He^3 + n + 3.27 \text{ MeV} \qquad [14.29(b)]$$

$$_1H^2 + _1H^2 \rightarrow {}_1H^3 + {}_1H^1 + 4.03 \text{ MeV} \qquad [14.29(c)]$$

In the first reaction, two protons combine to form a deuteron and a positron with a release of 0.42 MeV energy. In reaction [14.29(b)], two deuterons combine to form the light isotope of helium. In reaction (14.29(c)), two deuterons combine to form a triton and a proton. For fusion to take place, the two nuclei must come close enough so that attractive short-range nuclear force is able to affect them. However, since they are both positively charged particles, they experience coulomb repulsion. They, therefore, must have enough energy to overcome this coulomb barrier. The height of the barrier depends on the charges and radii of the two interacting nuclei. It can be shown, for example, that the barrier height for two protons is ~ 400 keV, and is higher for nuclei with higher charges. We can estimate the temperature at which two protons in a proton gas would (averagely) have enough energy to overcome the coulomb barrier :

$(3/2)k\,T = K$? 400 keV, which gives $T \sim 3 \times 10^9$ K.

When fusion is achieved by raising the temperature of the system so that particles have enough kinetic energy to overcome the coulomb repulsive behaviour, it is called *thermonuclear fusion.*

Thermonuclear fusion is the source of energy output in the interior of stars. The interior of the sun has a temperature of 1.5×10^7 K, which is considerably less than the estimated temperature required for fusion of particles of average energy. Clearly, fusion in the sun involves protons whose energies are much above the average energy.

The fusion reaction in the sun is a multi-step process in which the hydrogen is burned into helium. Thus, the fuel in the sun is the hydrogen in its core. The *proton-proton* (p, p) *cycle* by which this occurs is represented by the following sets of reactions:

$$_1H^1 + _1H^1 \rightarrow {}_1H^2 + e^+ + \nu + 0.42 \text{ MeV} \quad \text{(i)}$$

$$e^+ + e^- \rightarrow \gamma + \gamma + 1.02 \text{ MeV} \quad \text{(ii)}$$

$$_1H^2 + _1H^1 \rightarrow {}_2He^3 + \gamma + 5.49 \text{ MeV} \quad \text{(iii)}$$

$$_2He^3 + {}_2He^3 \rightarrow {}_2He^4 + {}_1H^1 + {}_1H^1 + 12.86 \text{ MeV} \quad \text{(iv)} \qquad (14.30)$$

For the fourth reaction to occur, the first three reactions must occur twice, in which case two light helium nuclei unite to form ordinary helium nucleus. If we consider the combination 2(i) + 2(ii) + 2(iii) +(iv), the net effect is

$$4\ {}_1H^1 + 2e^- \rightarrow {}_2He^4 + 2\nu + 6\gamma + 26.7 \text{ MeV}$$

or $$(4\ {}_1H^1 + 4e^-) \rightarrow ({}_2He^4 + 2e^-) + 2\nu + 6\gamma + 26.7 \text{ MeV} \qquad (14.31)$$

Thus, four hydrogen atoms combine to form an $_2He^4$ atom with a release of 26.7 MeV of energy.

Helium is not the only element that can be synthesized in the interior of a star. As the hydrogen in the core gets depleted and becomes helium, the core starts to cool. The star begins to collapse under its own gravity which increases the temperature of the core. If this temperature increases to about 10^8 K, fusion takes place again, this time of helium nuclei into carbon. This kind of process can generate through fusion higher and higher mass number elements. But elements more massive than those near the peak of the binding energy curve in Fig. 14.1 cannot be so produced.

The age of the sun is about 5×10^9 y and it is estimated that there is enough hydrogen in the sun to keep it going for another 5 billion years. After that, the hydrogen burning will stop and the sun will begin to cool and will start to collapse under gravity, which will raise the core temperature. The outer envelope of the sun will expand, turning it into the so called *red giant.*

NUCLEAR HOLOCAUST

In single uranium fission about 0.9×235 MeV (≈200 MeV) of energy is liberated. If each nucleus of about 50 kg of ^{235}U undergoes fission the amount of energy involved is about 4×10^{15}J. This energy is equivalent to about 20,000 tons of TNT, enough for a super explosion. Uncontrolled release of large nuclear energy is called an atomic explosion. On August 6, 1945 an atomic device was used in warfare for the first time. The US dropped an atom bomb on Hiroshima, Japan. The explosion was equivalent to 20,000 tons of TNT. Instantly the radioactive products devastated 10 sq km of the city which had 3,43,000 inhabitants. Of this number 66,000 were killed and 69,000 were injured; more than 67% of the city's structures were destroyed.

High temperature conditions for fusion reactions can be created by exploding a fission bomb. Super-explosions equivalent to 10 megatons of explosive power of TNT were tested in 1954. Such bombs which involve fusion of isotopes of hydrogen, deuterium and tritium are called hydrogen bombs. It is estimated that a nuclear arsenal sufficient to destroy every form of life on this planet several times over is in position to be triggered by the press of a button. Such a nuclear holocaust will not only destroy the life that exists now but its radioactive fallout will make this planet unfit for life for all times. Scenarios based on theoretical calculations predict a long *nuclear winter,* as the radioactive waste will hang like a cloud in the earth's atmosphere and will absorb the sun's radiation.

14.7.4 Controlled thermonuclear fusion

The natural thermonuclear fusion process in a star is replicated in a thermonuclear fusion device. In controlled fusion reactors, the aim is to generate steady power by heating the nuclear fuel to a temperature in the range of 10^8 K. At these temperatures, the fuel is a mixture of positive ions and electrons (plasma). The challenge is to confine this plasma, since no container can stand such a high temperature. Several countries around the world including India are developing techniques in this connection. If successful, fusion reactors will hopefully supply almost unlimited power to humanity.

SUMMARY

1. An atom has a nucleus. The nucleus is positively charged. The radius of the nucleus is smaller than the radius of an atom by a factor of 10^4. More than 99.9% mass of the atom is concentrated in the nucleus.
2. On the atomic scale, mass is measured in atomic mass units (u). By definition, 1 atomic mass unit (1u) is $1/12^{th}$ mass of one atom of ^{12}C;

   $1u = 1.660563 \times 10^{-27}$ kg.
3. A nucleus contains a neutral particle called neutron. Its mass is almost the same as that of proton
4. The atomic number Z is the number of protons in the atomic nucleus of an element. The mass number A is the total number of protons and neutrons in the atomic nucleus; $A = Z+N$; Here Ndenotes the number of neutrons in the nucleus.

   A nuclear species or a nuclide is represented as ${}_ZX^A$, where ${}_ZX^A$ is the chemical symbol of the species.

   Nuclides with the same atomic number Z, but different neutron number N are called *isotopes*. Nuclides with the same A are *isobars* and those with the same N are *isotones*.

   Most elements are mixtures of two or more isotopes. The atomic mass of an element is a weighted average of the masses of its isotopes. The masses are the relative abundances of the isotopes.
5. A nucleus can be considered to be spherical in shape and assigned a radius. Electron scattering experiments allow determination of the nuclear radius; it is found that radii of nuclei fit the formula

   $R = R\,A^{1/3}$,

   whereR_0 = a constant = 1.2 fm. This implies that the nuclear density is independent of A. It is of the order of 10^{17} kg/m^3.
6. Neutrons and protons are bound in a nucleus by the short-range strong nuclear force. The nuclear force does not distinguish between neutron and proton.
7. The nuclear mass M is always less than the total mass, Σm, of its constituents. The difference in mass of a nucleus and its constituents is called the *mass defect*,

   $\Delta M = (Z\,m_p + (A - Z)m_n) - M$

   Using Einstein's mass energy relation, we express this mass difference in terms of energy as

   $\Delta E_b = \Delta M\,c^2$

   The energy ΔE_brepresents the *binding energy* of the nucleus. In the mass number range A = 30 to 170, the binding energy per nucleon is nearly constant, about 8 MeV/nucleon.
8. Energies associated with nuclear processes are about a million times larger than chemical process.
9. The Q-value of a nuclear process is

   Q = final kinetic energy – initial kinetic energy.

   Due to conservation of mass-energy, this is also,

   Q = (sum of initial masses – sum of final masses)c^2
10. Radioactivity is the phenomenon in which nuclei of a given species transform by giving out α or βorγrays; α-rays are helium nuclei; β-rays are electrons. γ-rays are electromagnetic radiation of wavelengths shorter than X-rays;

11. Law of radioactive decay :$N(t) = N(0)\, e^{-\lambda t}$

    where λ is the decay constant or disintegration constant.

    The half-life $T_{1/2}$ of a radionuclide is the time in which N has been reduced to one-half of its initial value. The mean life τ is the time at which N has been reduced to e^{-1} of its initial value

    $$T_{1/2} = \frac{\ln 2}{\lambda} = \tau \ln 2$$

12. Energy is released when less tightly bound nuclei are transmuted into more tightly bound nuclei. In fission, a heavy nucleus like ${}_{92}U^{235}$ breaks into two smaller fragments, e.g., ${}_{92}U^{235} + {}_{0}n^{1} \rightarrow {}_{51}Sb^{133} + {}_{41}Nb^{99} + 4\ {}_{0}n^{1}$.

13. The fact that more neutrons are produced in fission than are consumed gives the possibility of a chain reaction with each neutron that is produced triggering another fission. The chain reaction is uncontrolled and rapid in a nuclear bomb explosion. It is controlled and steady in a nuclear reactor. In a reactor, the value of the neutron multiplication factor k is maintained at 1.

14. In fusion, lighter nuclei combine to form a larger nucleus. Fusion of hydrogen nuclei into helium nuclei is the source of energy of all stars including our sun.

Physical Quantity	**Symbol**	**Dimensions**	**Units**	**Remarks**
Atomic mass unit		[M]	u	Unit of mass for expressing atomic or nuclear masses. One atomic mass unit equals 1/12th of the mass of 12C atom.
Disintegration or decay constant	λ	[T –1]	s^{-1}	
Half-life	$T_{1/2}$	[T]	s	Time taken for the decay of one-half ofthe initialnumber of nuclei present in a radioactive sample.
Mean life	τ	[T]	s	Time at which number ofnuclei has been reduced to e^{-1} of its initial value
Activity of a radio-active sample	R	$[T^{-1}]$	Bq	Measure of the activity of a radioactive source.

VERY SHORT ANSWER QUESTIONS

1. What are isotopes and isobars?
2. What is a.m.u.? What is its equivalent energy?
3. What will be the ratio of the radii of two nuclei of mass numbers A_1 and A_2?
4. Natural radioactive nuclei are mostly nuclei of high mass number. Why?
5. A nucleus contains no electrons but can emit them. How?
6. What are the units and dimensions of disintegration constant?
7. Why do all electrons emitted during β-decay not have the same energy?
8. Neutrons cannot produce ionization. Why?
9. What are delayed neutrons?
10. What are thermal neutrons? What is their importance?
11. What is the role of controlling rods in a nuclear reactor?
12. Define Becqueral and Curie.
13. What is the function of moderator in a nuclear reactor?

SHORT ANSWER QUESTIONS

1. Write a short note on the discovery of neutron.
2. What are the properties of neutron?
3. What are nuclear forces? Write their properties.
4. Define half life period and decay constant for a radioactive substance. Deduce relation between them.
5. What is nuclear fission? Give an example to illustrate it.
6. What is nuclear fusion? Write the conditions for nuclear fusion to occur.
7. Distinguish between nuclear fission and nuclear fusion.

LONG ANSWER QUESTIONS

1. Define mass defect and binding energy. How does binding energy per nucleon vary with mass number? What is its significance?
2. What is radioactivity? State the law of radioactive decay. Show that radioactive decay is exponential in nature.
3. Explain the principle and working of a nuclear reactor with the help of a labelled diagram.
4. Explain the source of stellar energy. Explain the carbon-nitrogen cycle, proton-proton cycle occurring in stars.

CHAPTER 15

SEMICONDUCTOR ELECTRONICS: MATERIALS, DEVICES AND SIMPLE CIRCUITS

15.1 INTRODUCTION

Devices in which a controlled flow of electrons can be obtained are the basic *building blocks* of all the electronic circuits. Before the discovery of transistor in 1948, such devices were mostly vacuum tubes (also called valves) like the vacuum diode which has two electrodes, viz., anode (often called plate) and cathode; triode which has three electrodes – cathode, plate and grid; tetrode and pentode (respectively with 4 and 5 electrodes). In a vacuum tube, the electrons are supplied by a heated cathode and the controlled flow of these electrons *in vacuum* is obtained by varying the voltage between its different electrodes. Vacuum is required in the inter-electrode space; otherwise the moving electrons may lose their energy on collision with the air molecules in their path. In these devices the electrons can flow only from the cathode to the anode (i.e., only in one direction). Therefore, such devices are generally referred to as *valves*. These vacuum tube devices are bulky, consume high power, operate generally at high voltages (~100 V) and have limited life and low reliability. The seed of the development of modern *solid-state semiconductor electronics* goes back to 1930's when it was realised that some solid-state semiconductors and their junctions offer the possibility of controlling the number and the direction of flow of charge carriers through them. Simple excitations like light, heat or small applied voltage can change the number of mobile charges in a semiconductor. Note that the supply and flow of charge carriers in the semiconductor devices are *within the solid itself*, while in the earlier vacuum tubes/valves, the mobile electrons were obtained from a heated cathode and they were made to flow in an *evacuated* space or vacuum. No external heating or large evacuated space is required by the semiconductor devices. They are small in size, consume low power, operate at low voltages and have long life and high reliability. Even the Cathode Ray Tubes (CRT) used in television and computer monitors which work on the principle of vacuum tubes are being replaced by Liquid Crystal Display (LCD) monitors with supporting solid state electronics. Much before the full implications of the semiconductor devices was formally understood, a naturally occurring crystal of *galena* (Lead sulphide, PbS) with a metal point contact attached to it was used as *detector* of radio waves.

15.2 CLASSIFICATION OF METALS, CONDUCTORS AND SEMICONDUCTORS

On the basis of conductivity

On the basis of the relative values of electrical conductivity (σ) or resistivity ($\rho = 1/\sigma$), the solids are broadly classified as:

(i) ***Metals:*** They possess very low resistivity (or high conductivity).

$\rho \sim 10^{-2} - 10^{-8}\ \Omega$ m

$\sigma \sim 10^{2} - 10^{8}$ S m^{-1}

(ii) ***Semiconductors:*** They have resistivity or conductivity intermediate to metals and insulators.

$\rho \sim 10^{-5} - 10^{6}\ \Omega$ m

$\sigma \sim 10^{5} - 10^{-6}$ S m^{-1}

(iii) ***Insulators:*** They have high resistivity (or low conductivity).

$\rho \sim 10^{11} - 10^{19}\ \Omega$ m

$\sigma \sim 10^{-11} - 10^{-19}$ S m^{-1}

The values of ρ and σ given above are indicative of magnitude and could well go outside the ranges as well. Relative values of the resistivity are not the only criteria for distinguishing metals, insulators and semiconductors from each other. There are some other differences, which will become clear as we go along in this chapter.

Our interest in this chapter is in the study of semiconductors which could be:

i) *Elemental semiconductors*: Si and Ge

ii) Compound semiconductors: Examples are

- Inorganic: CdS, GaAs, CdSe, InP, etc
- Organic: anthracene, doped pthalocyanines, etc.
- Organic polymers: polypyrrole, polyaniline, polythiophene, etc.

Most of the currently available semiconductor devices are based on elemental semiconductors Si or Ge and compound *inorganic* semiconductors. However, after 1990, a few semiconductor devices using organic semiconductors and semiconducting polymers have been developed signaling the birth of a futuristic technology of polymer-electronics and molecular-electronics. In this chapter, we will restrict ourselves to the study of inorganic semiconductors, particularly elemental semiconductors Si and Ge. The general concepts introduced here for discussing the elemental semiconductors, by-and-large, apply to most of the compound semiconductors as well.

On the basis of energy bands

According to the Bohr atomic model, in an *isolated atom* the energy of any of its electrons is decided by the orbit in which it revolves. But when the atoms come together to form a solid they are close to each other. So the outer orbits of electrons from neighbouring atoms would come very close or could even overlap. This would make the nature of electron motion in a solid very different from that in an isolated atom.

Inside the crystal each electron has a unique position and no two electrons see exactly the same pattern of surrounding charges. Because of this, each electron will have a different *energy level*. These different energy levels with continuous energy variation form what are called *energy bands*. The energy band which includes the energy levels of the valence electrons is called the *valence band*. The energy band above the valence band is called the *conduction band*. With no external energy, all the valence electrons will reside in the valence band. If the lowest level in the conduction band happens to be lower than the highest level of the valence band, the electrons from the valence band can easily move into the conduction band. Normally the conduction band is empty. But when it overlaps on the valence band electrons can move freely into it. This is the case with metallic conductors.

If there is some gap between the conduction band and the valence band, electrons in the valence band all remain bound and no free electrons are available in the conduction band. This makes the material an insulator. But some of the electrons from the valence band may gain external energy to cross the gap between the conduction band and the valence band. Then these electrons will move into the conduction band. At the same time they will create vacant energy

levels in the valence band where other valence electrons can move. Thus the process creates the possibility of conduction due to electrons in conduction band as well as due to vacancies in the valence band.

Let us consider what happens in the case of Si or Ge crystal containing *N* atoms. For Si, the outermost orbit is the third orbit (n = 3), while for Ge it is the fourth orbit (n = 4). The number of electrons in the outermost orbit is 4 (2*s* and 2*p* electrons). Hence, the total number of outer electrons in the crystal is 4*N*. The maximum possible number of electrons in the outer orbit is 8 (2*s* + 6*p* electrons). So, for the 4*N* valence electrons there are 8*N* available energy states. These 8*N* discrete energy levels can either form a continuous band or they may be grouped in different bands depending upon the distance between the atoms in the crystal (see box on Band Theory of Solids).

At the distance between the atoms in the crystal lattices of Si and Ge, the energy band of these 8*N* states is split apart into two which are separated by an *energy gap* E_g (Fig. 15.1). The lower band which is completely occupied by the 4*N* valence electrons at temperature of absolute zero is the *valence band*. The other band consisting of 4*N* energy states, called the *conduction band*, is completely empty at absolute zero.

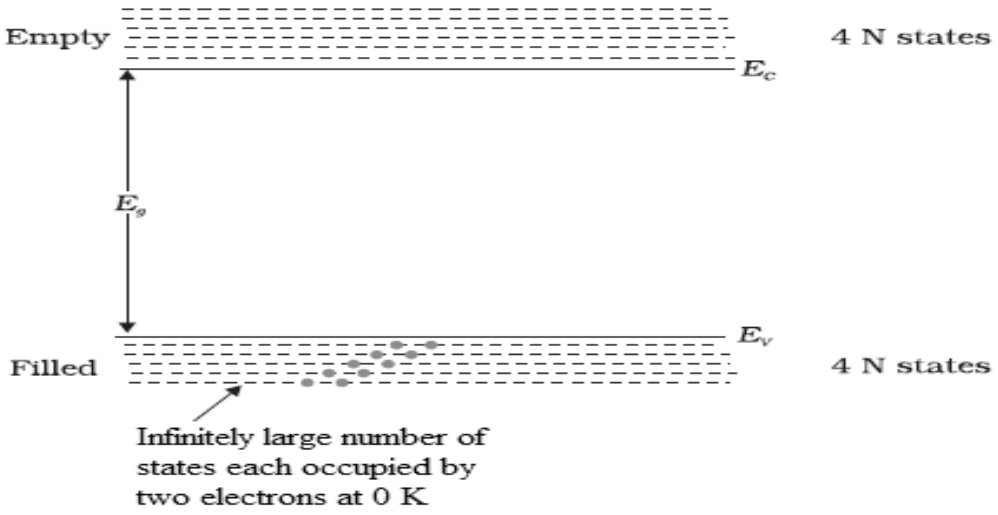
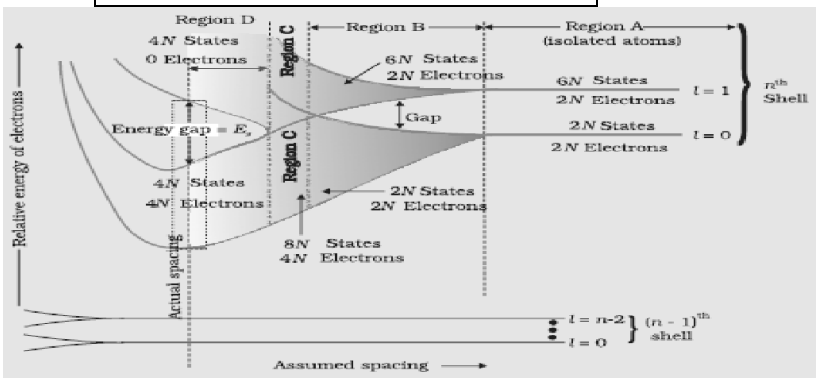


Fig. 15.1 The energy band positions in a semiconductor at 0 K. The upper band, called the conduction band, consists of infinitely large number of closely spaced energy states. The lower band, called the valence band, consists of closely spaced completely filled energy states.

The lowest energy level in the conduction band is shown as E_C and highest energy level in the valence band is shown as E_V. Above E_C and below E_V there are a large number of closely spaced energy levels, as shown in Fig. 15.1.

The gap between the top of the valence band and bottom of the conduction band is called the *energy band gap* (Energy gap E_g). It may be large, small, or zero, depending upon the material. These different situations are depicted in Fig. 15.2 and discussed:

BAND THEORY OF SOLIDS



Consider that the Si or Ge crystal contains N atoms. Electrons of each atom will have discrete energies in different orbits. The electron energy will be same if all the atoms are isolated, i.e., separated from each other by a large distance. However, in a crystal, the atoms are close to each other (2 to 3 Å) and therefore the electrons interact with each other and also with the neighbouring atomic cores. The overlap (or interaction) will be more felt by the electrons in the outermost orbit while the inner orbit or core electron energies may remain unaffected. Therefore, for understanding electron energies in Si or Ge crystal, we need to consider the changes in the energies of the electrons in the outermost orbit only. For Si, the outermost orbit is the third orbit (n = 3), while for Ge it is the fourth orbit (n = 4). The number of electrons in the outermost orbit is 4 (2s and 2p electrons). Hence, the total number of outer electrons in the crystal is 4N. The maximum possible number of outer electrons in the orbit is 8 (2s + 6p electrons). So, out of the 4N electrons, 2N electrons are in the 2N s-states (orbital quantum number l = 0) and 2N electrons are in the available 6N p-states. Obviously, some p-electron states are empty as shown in the extreme right of Figure. This is the case of well separated or isolated atoms [region A of Figure]. Suppose these atoms start coming nearer to each other to form a solid. The energies of these electrons in the outermost orbit may change (both increase and decrease) due to the interaction between the electrons of different atoms. The 6N states for l = 1, which originally had identical energies in the isolated atoms, spread out and form an energy band [region B in Figure]. Similarly, the 2N states for l = 0, having identical energies in the isolated atoms, split into a second band (carefully see the region B of Figure) separated from the first one by an energy gap. At still smaller spacing, however, there comes a region in which the bands merge with each other. The lowest energy state that is a split from the upper atomic level appears to drop below the upper state that has come from the lower atomic level. In this region (region C in Figure), no energy gap exists where the upper and lower energy states get mixed. Finally, if the distance between the atoms further decreases, the energy bands again split apart and are separated by an energy gap Eg (region D in Figure). The total number of available energy states 8N has been re-apportioned between the two bands (4N states each in the lower and upper energy bands). Here the significant point is that there are exactly as many states in the lower band (4N) as there are available valence electrons from the atoms (4N). Therefore, this band (called the valence band) is completely filled while the upper band is completely empty. The upper band is called the conduction band.

Case I: This refers to a situation, as shown in Fig. 15.2(a). One can have a metal either when the conduction band is partially filled or the balanced band is partially empty or when the conduction and valance bands overlap. When there is overlap electrons from valence band can easily move into the conduction band. This situation makes a large number of electrons available for electrical conduction. When the valence band is partially empty, electrons from its lower level can move to higher level making conduction possible. Therefore, the resistance of such materials is low or the conductivity is high.

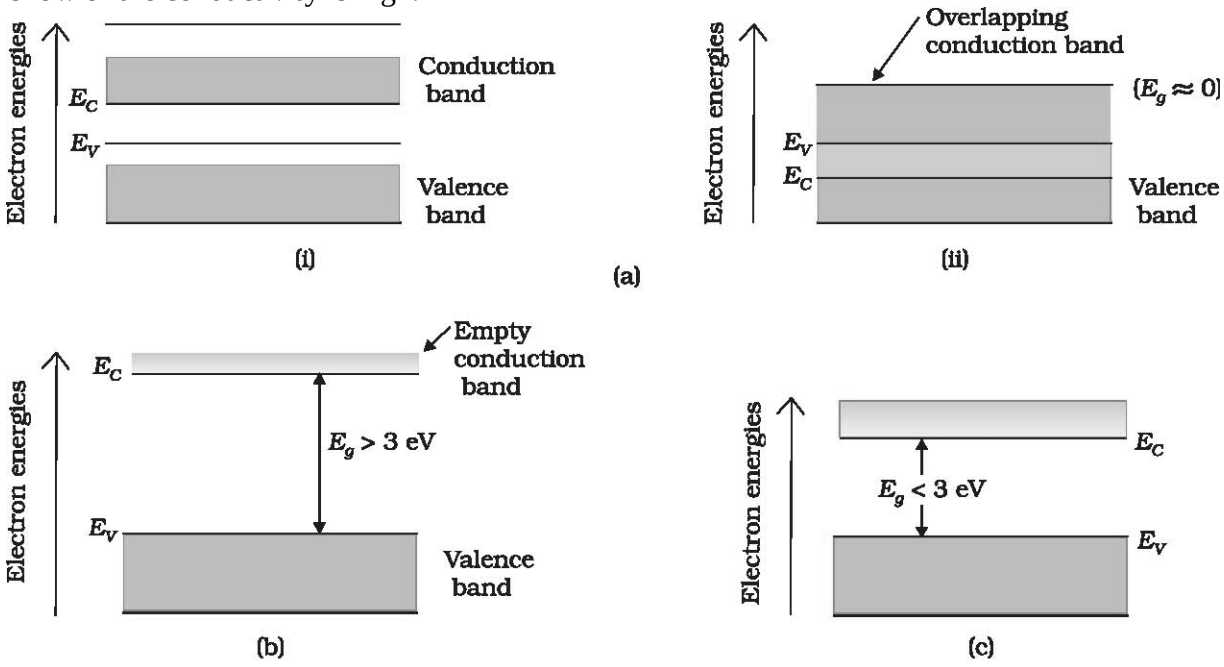


Fig. 15.2 Difference between energy levels of (a) metals (b) insulators and (c) semiconductors

Case II: In this case, as shown in Fig. 15.2(b), a large band gap E_g exists ($E_g > 3$ eV). There are no electrons in the conduction band, and therefore no electrical conduction is possible. Note that the energy gap is so large that electrons cannot be excited from the valence band to the conduction band by thermal excitation. This is the case of *insulators*.

Case III: This situation is shown in Fig. 15.2 (c). Here a finite but small band gap ($E_g < 3$ eV) exists. Because of the small band gap, at room temperature some electrons from valence band can acquire enough energy to cross the energy gap and enter the *conduction band*. These electrons (though small in numbers) can move in the conduction band. Hence, the resistance of *semiconductors* is not as high as that of the insulators.

In this section we have made a broad classification of metals, conductors and semiconductors. In the section which follows you will learn the conduction process in semiconductors.

15.3 INTRINSIC SEMICONDUCTOR

We shall take the most common case of Ge and Si whose lattice structure is shown in Fig. 15.3. These structures are called the diamond-like structures. Each atom is surrounded by four nearest neighbours. We know that Si and Ge have four valence electrons. In its crystalline

structure, every Si or Ge atom tends *to share* one of its four valence electrons with each of its four nearest neighbour atoms, and also *to take share* of one electron from each such neighbour. These shared electron pairs are referred to as forming a *covalent bond* or simply a *valence bond*. The two shared electrons can be assumed to shuttle back-and forth between the associated atoms holding them together strongly. Figure 15.4 schematically shows the 2-dimensional representation of Si or Ge structure shown in Fig. 15.3 which over emphasises the covalent bond. It shows an idealised picture in which no bonds are broken (all bonds are intact). Such a situation arises at low temperatures. As the temperature increases, more thermal energy becomes available to these electrons and some of these electrons may break–away (becoming *free* electrons contributing to conduction). The thermal energy effectively ionises only a few atoms in the crystalline lattice and creates a *vacancy* in the bond as shown in Fig. 15.5(a). The neighborhood, from which the free electron (with charge $-q$) has come out leaves a vacancy with an effective charge ($+q$). This *vacancy* with the effective positive electronic charge is called a *hole*. The hole behaves as an *apparent free particle* with effective positive charge.

In intrinsic semiconductors, the number of free electrons, n_e is equal to the number of holes, n_h. That is

$$n_e = n_h = n_i \qquad (15.1)$$

where n_i is called intrinsic carrier concentration. Semiconductors posses the unique property in which, apart from electrons, the holes also move.

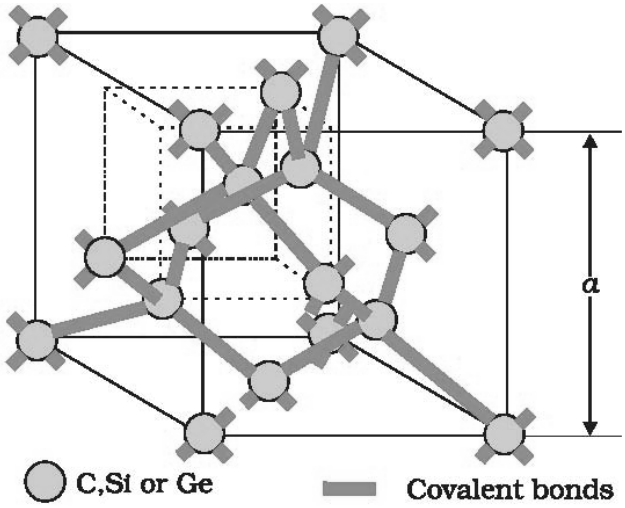


Fig. 15.3 Three-dimensional diamonds- like crystal structure for Carbon, Silicon or Germanium with respective lattice spacing a equal to 3.56, 5.43 and 5.66 Å.

Suppose there is a hole at site 1 as shown in Fig. 14.5(a). The movement of holes can be visualised as shown in Fig. 14.5(b). An electron from the covalent bond at site 2 may jump to the vacant site 1 (hole). Thus, after such a jump, the hole is at site 2 and the site 1 has now an electron. Therefore, apparently, the hole has moved from site 1 to site 2. Note that the electron originally set free [Fig. 14.5(a)] is not involved in this process of hole motion. The free electron moves completely independently as conduction electron and gives rise to an electron current, I_e under an applied electric field. Remember that the motion of hole is only a convenient way of describing the actual motion of *bound* electrons, whenever there is an empty bond anywhere in

the crystal. Under the action of an electric field, these holes move towards negative potential giving the hole current, I_h. The total current, I is thus the sum of the electron current I_e and the hole current I_h

$$I = I_e + I_h \qquad (15.2)$$

It may be noted that apart from the *process of generation* of conduction electrons and holes, a simultaneous *process of recombination* occur in which the electrons *recombine* with the holes. At equilibrium, the rate of generation is equal to the rate of recombination of charge carriers. The recombination occurs due to an electron colliding with a hole.

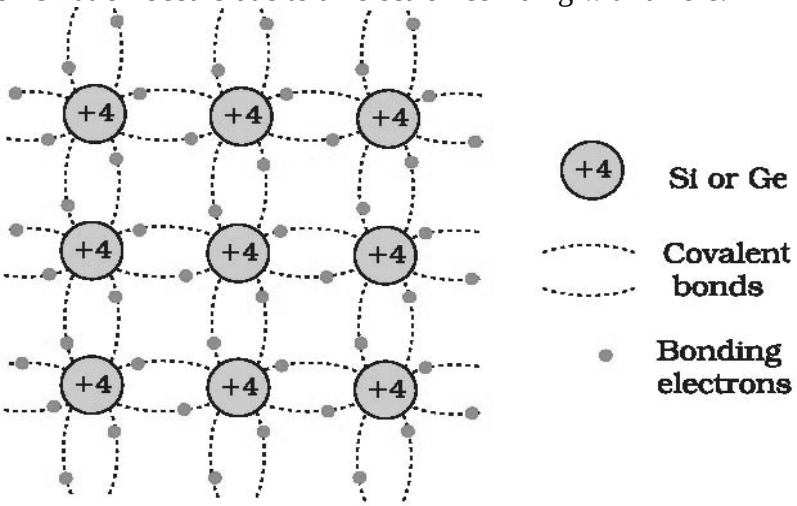


Fig.15.4 Schematic two-dimensional representation of Si or Ge structure showing covalent bonds at low temperature (all bonds intact).+4 symbol indicates inner cores of Si or Ge.

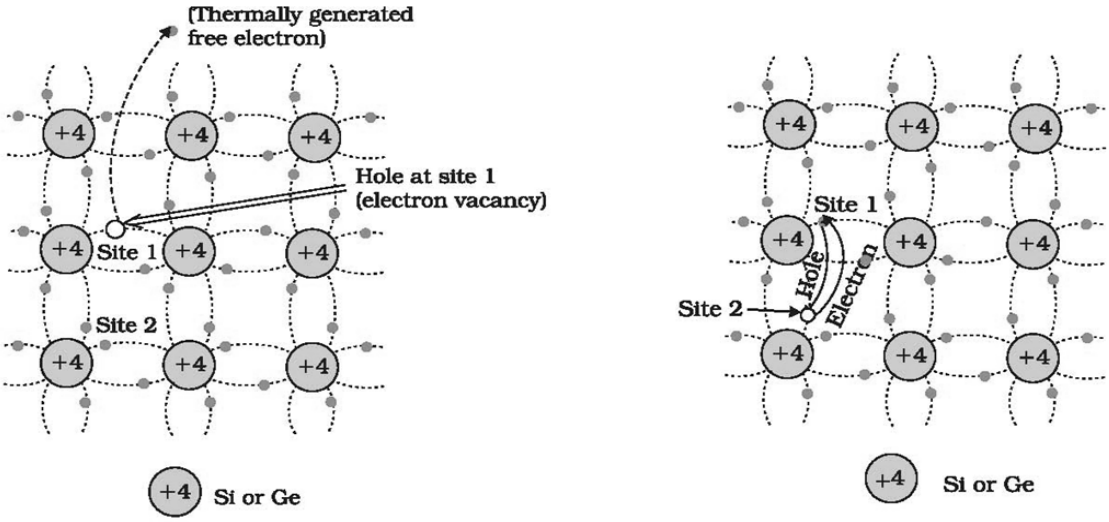


Fig. 15.5 (a) Schematic model of generation of hole at site 1 and conduction electron due to thermal energy at moderate temperatures. (b) Simplified representation of possible thermal motion of a hole. The electron from the lower left hand covalent bond (site 2)

goes to the earlier hole site1, leaving a hole at its site indicating an apparent movement of the hole from site 1 to site 2.

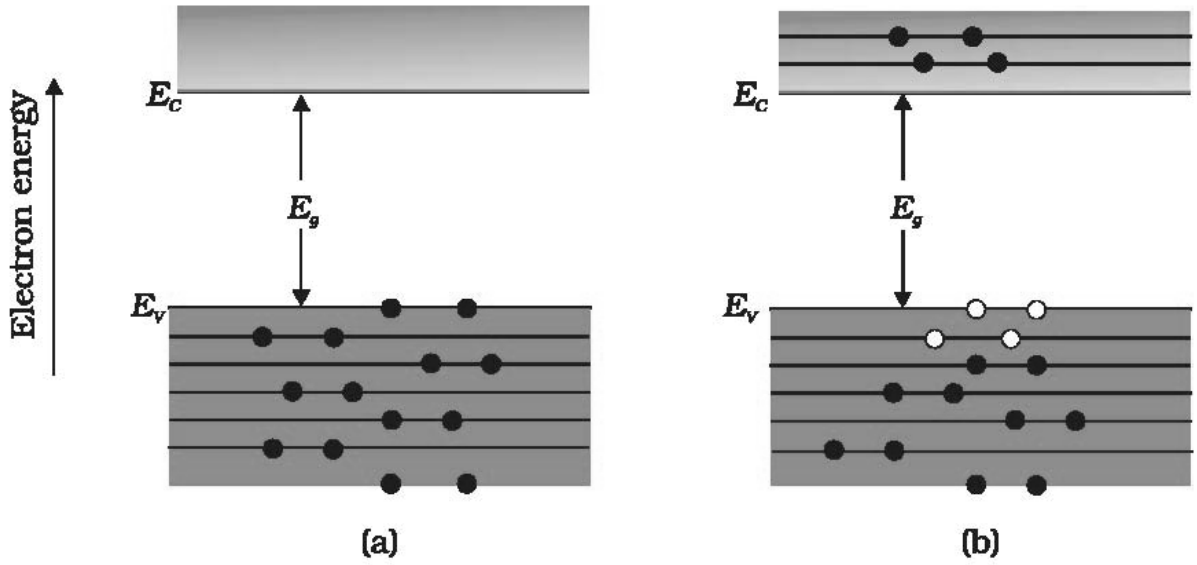


Fig. 15.6 (a) An intrinsic semiconductor at T = 0 K behaves like insulator. (b) At T > 0 K, four thermally generated electron-hole pairs. The filled circles (●) represent electrons and empty fields ( ) represent holes.

An intrinsic semiconductor will behave like an insulator at T = 0 K as shown in Fig. 15.6 (a). It is the thermal energy at higher temperatures (T > 0K), which excites some electrons from the valence band to the conduction band. These thermally excited electrons at T > 0K, partially occupy the conduction band. Therefore, the energy-band diagram of an intrinsic semiconductor will be as shown in Fig. 15.6(b). Here, some electrons are shown in the conduction band. These have come from the valence band leaving equal number of holes there.

15.4 EXTRINSIC SEMICONDUCTOR

The conductivity of an intrinsic semiconductor depends on its temperature, but at room temperature its conductivity is very low. As such, no important electronic devices can be developed using these semiconductors. Hence there is a necessity of improving their conductivity. This can be done by making use of impurities.

When a small amount, say, a few parts per million (ppm), of a suitable impurity is added to the pure semiconductor, the conductivity of the semiconductor is increased manifold. Such materials are known as e*xtrinsic semiconductors* or *impurity semiconductors.* The deliberate addition of a desirable impurity is called *doping* and the impurity atoms are called *dopants*. Such a material is also called a *doped semiconductor*. The dopant has to be such that it does not distort the original pure semiconductor lattice. It occupies only a very few of the original semiconductor atom sites in the crystal. A necessary condition to attain this is that the sizes of the dopant and the semiconductor atoms should be nearly the same.

There are two types of dopants used in doping the tetravalent Si or Ge:

(i) Pentavalent (valency 5); like Arsenic (As), Antimony (Sb), Phosphorous(P), etc.

(ii) Trivalent (valency 3); like Indium (In), Boron (B), Aluminium (Al), etc.

We shall now discuss how the doping changes the number of charge carriers (and hence the conductivity) of semiconductors. Si or Ge belongs to the fourth group in the Periodic table and, therefore, we choose the dopant element from nearby fifth or third group, expecting and taking care that the size of the dopant atom is nearly the same as that of Si or Ge. Interestingly, the pentavalent and trivalent dopants in Si or Ge give two entirely different types of semiconductors as discussed below.

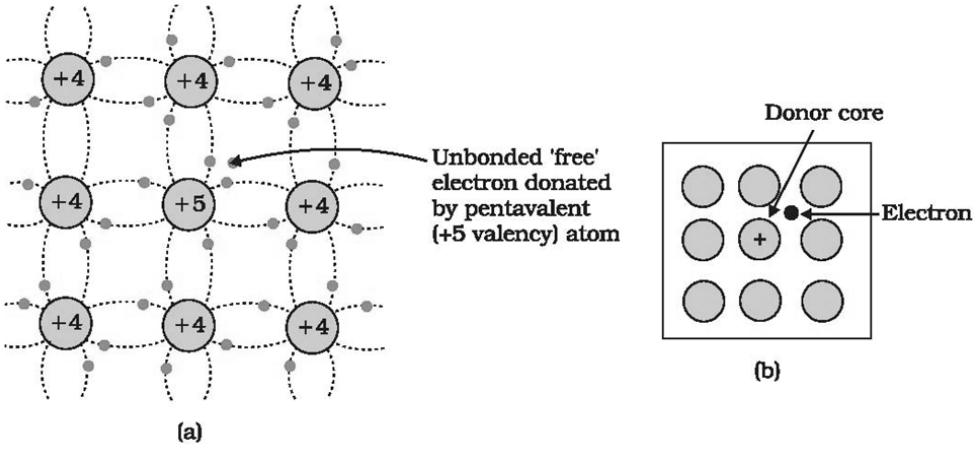


Fig. 15.7 (a) Pentavalent donor atom (As, Sb, P, etc.) doped for tetravalent Si or Ge giving n-type semiconductor, and (b) Commonly used schematic representation of n-type material which shows only the fixed cores of the substituent donors with one additional effective positive charge and its associated extra electron.

(i) n-type semiconductor

Suppose we dope Si or Ge with a pentavalent element as shown in Fig. 15.7. When an atom of +5 valency element occupies the position of an atom in the crystal lattice of Si, four of its electrons bond with the four silicon neighbours while the fifth remains very weakly bound to its parent atom. This is because the four electrons participating in bonding are seen as part of the effective core of the atom by the fifth electron. As a result the ionisation energy required to set this electron free is very small and even at room temperature it will be free to move in the lattice of the semiconductor. For example, the energy required is ~ 0.01 eV for germanium, and 0.05 eV for silicon, to separate this electron from its atom. This is in contrast to the energy required to jump the forbidden band (about 0.72 eV for germanium and about 1.1 eV for silicon) at room temperature in the intrinsic semiconductor. Thus, the pentavalent dopant is donating one extra electron for conduction and hence is known as *donor* impurity. The number of electrons made available for conduction by dopant atoms depends strongly upon the doping level and is independent of any increase in ambient temperature. On the other hand, the number of free electrons (with an equal number of holes) generated by Si atoms, increases weakly with temperature. In a doped semiconductor the total number of conduction electrons

n_e is due to the electrons contributed by donors and those generated intrinsically, while the total number of holes n_h is only due to the holes from the intrinsic source. But the rate of recombination of holes would increase due to the increase in the number of electrons. As a result, the number of holes would get reduced further.

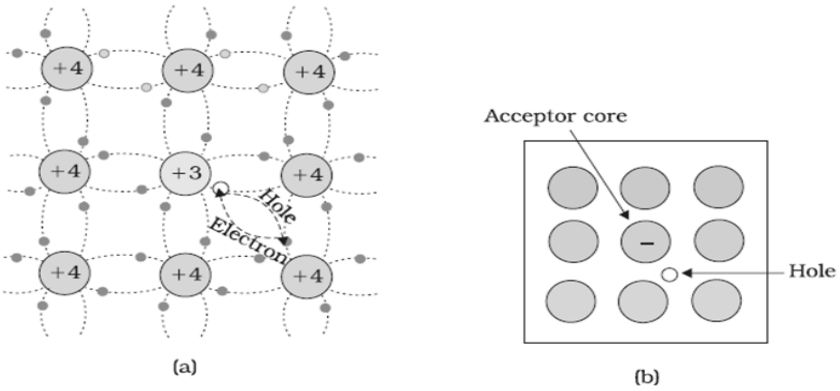


Fig. 15.8 (a) Trivalent acceptor atom (In, Al, B etc.) doped in tetravalent Si or Ge lattice giving p-type semiconductor. (b) Commonly used schematic representation of p-type material which shows only the fixed core of the substituent acceptor with one effective additional negative charge and its associated hole.

Thus, with proper level of doping the number of conduction electrons can be made much larger than the number of holes. Hence in an extrinsic semiconductor doped with pentavalent impurity, electrons become the *majority carriers* and holes the *minority carriers*.These semiconductors are, therefore, known as n-*type semiconductors*. For n-type semiconductors, we have,

$$n_e >> n_h \qquad (15.3)$$

ii) p-type semiconductor

This is obtained when Si or Ge is doped with a trivalent impurity like Al, B, In, etc. The dopant has one valence electron less than Si or Ge and, therefore, this atom can form covalent bonds with neighbouring three Si atoms but does not have any electron to offer to the fourth Si atom. So the bond between the fourth neighbour and the trivalent atom has a vacancy or hole as shown in Fig. 15.8. Since the neighbouring Si atom in the lattice wants an electron in place of a hole, an electron in the outer orbit of an atom in the neighbourhood may jump to fill this vacancy, leaving a vacancy or hole at its own site. Thus the *hole* is available for conduction. Note that the trivalent foreign atom becomes effectively negatively charged when it shares fourth electron with neighbouring Si atom. Therefore, the dopant atom of p-type material can be treated as *core of one negative charge* along with its associated hole as shown in Fig. 15.8 (b). It is obvious that one *acceptor* atom gives one *hole*. These holes are in addition to the intrinsically generated holes while the source of conduction electrons is only intrinsic generation. Thus, for such a material, the holes are the majority carriers and electrons are minority carriers. Therefore, extrinsic semiconductors doped with trivalent impurity are called p-*type semiconductors*. For p-type semiconductors, the recombination process will reduce the number (n_i)of intrinsically generated electrons to n_e. We have, for p-type semiconductors

$$n_h >> n_e \qquad (15.4)$$

Note that *the crystal maintains an overall charge neutrality as the charge of additional charge carriers is just equal and opposite to that of the ionised cores in the lattice.*

In extrinsic semiconductors, because of the abundance of majority current carriers, the minority carriers produced thermally have more chance of meeting majority carriers and thus getting destroyed. Hence, the dopant, by adding a large number of current carriers of one type, which become the majority carriers, indirectly helps to reduce the intrinsic concentration of minority carriers.

The semiconductor's energy band structure is affected by doping. In the case of extrinsic semiconductors, additional energy states due to donor impurities (E_D) and acceptor impurities (E_A) also exist. In the energy band diagram of n-type Si semiconductor, the donor energy level E_D is slightly below the bottom E_C of the conduction band and electrons from this level move into the conduction band with very small supply of energy. At room temperature, most of the donor atoms get ionised but very few ($\sim 10^{-12}$) atoms of Si get ionised. So the conduction band will have most electrons coming from the donor impurities, as shown in Fig. 15.9(a). Similarly, for p-type semiconductor, the acceptor energy level E_A is slightly above the top E_V of the valence band as shown in Fig. 15.9(b). With very small supply of energy an electron from the valence band can jump to the level E_A and ionise the acceptor negatively. (Alternately, we can also say that with very small supply of energy the hole from level E_A sinks down into the valence band. Electrons rise up and holes fall down when they gain external energy.) At room temperature, most of the acceptor atoms get ionised leaving holes in the valence band. Thus at room temperature the density of holes in the valence band is predominantly due to impurity in the extrinsic semiconductor.

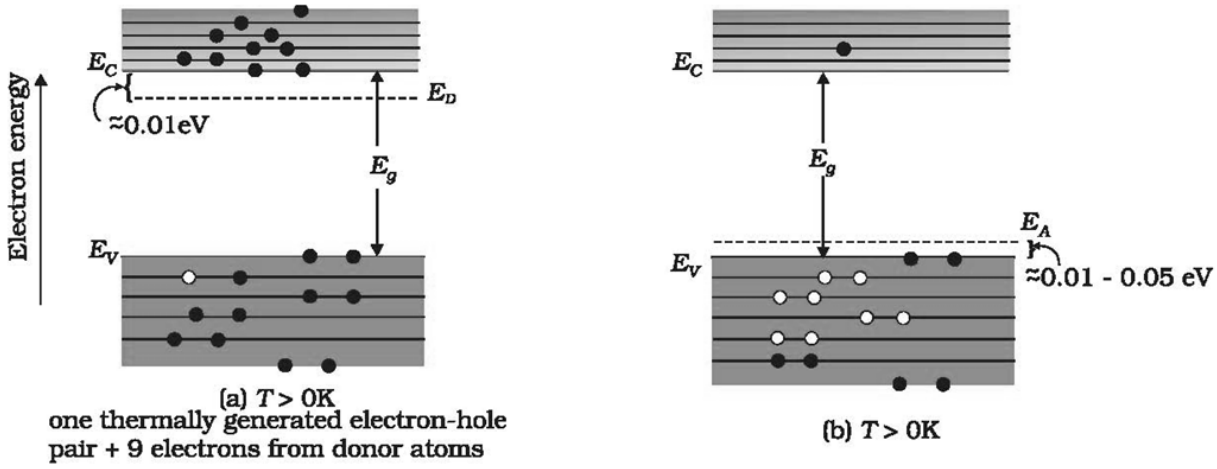


Fig. 15.9 Energy bands of (a) n-type semiconductor at T > 0K (b) p-type semiconductor at T > 0K.

The electron and hole concentration in a semiconductor *in thermal equilibrium* is given by

$$n_e n_h = n_i^2 \qquad (15.5)$$

Though the above description is grossly approximate and hypothetical, it helps in understanding the difference between metals, insulators and semiconductors (extrinsic and intrinsic) in a simple manner. The difference in the resistivity of C, Si and Ge depends upon the energy gap between their conduction and valence bands. For C (diamond), Si and Ge, the energy gaps are 5.4 eV, 1.1 eV and 0.7 eV, respectively. Sn also is a group IV element but it is a metal because the energy gap in its case is 0 eV.

15.5 p-n JUNCTION

A p-n junction is the basic building block of many semiconductor devices like diodes, transistor,etc. A clear understanding of the junction behaviour is important to analyse the working of other semiconductor devices. We will now try to understand how a junction is formed and how the junction behaves under the influence of external applied voltage (also called *bias*).

15.5.1 p-n junction formation

Consider a thin p-type silicon (p-Si) semiconductor wafer. By adding precisely a small quantity of pentavelent impurity, part of the p-Si wafer can be converted into n-Si. There are several processes by which a semiconductor can be formed. The wafer now contains p-region and n-region and a metallurgical junction between p-, and n- region.

Two important processes occur during the formation of a p-n junction: *diffusion* and *drift*. We know that in an n-type semiconductor, the concentration of electrons (number of electrons per unit volume) is more compared to the concentration of holes. Similarly, in a p-type semiconductor, the concentration of holes is more than the concentration of electrons. During the formation of p-n junction, and due to the concentration gradient across p-, and n- sides, holes diffuse from p-side to n-side (p → n) and electrons diffuse from n-side to p-side (n → p). This motion of charge carries gives rise to diffusion current across the junction.

When an electron diffuses from n → p, it leaves behind an ionised donor on n-side. This ionised donor (positive charge) is immobile as it is bonded to the surrounding atoms. As the electrons continue to diffuse from n → p, a layer of positive charge (or positive space-charge region) on n-side of the junction is developed.

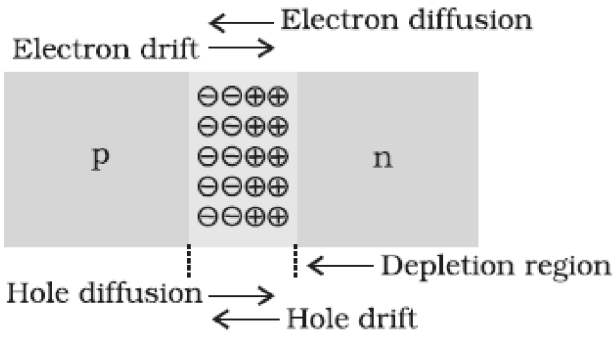


Fig. 15.10 p-n junction formation process.

Similarly, when a hole diffuses from p → n due to the concentration gradient, it leaves behind an ionised acceptor (negative charge) which is immobile. As the holes continue to diffuse, a layer of negative charge (or negative space-charge region) on the p-side of the junction is developed. This space-charge region on either side of the junction together is known as *depletion region* as the electrons and holes taking part in the initial movement across the junction *depleted* the region of its free charges (Fig. 15.10). The thickness of depletion region is of the order of one-tenth of a micrometer. Due to the positive space-charge region on n-side of the junction and negative space charge region on p-side of the junction, an electric field directed from positive charge towards negative charge develops. Due to this field, an electron on p-side of

the junction moves to n-side and a hole on n-side of the junction moves to p-side. The motion of charge carriers due to the electric field is called drift. Thus a drift current, which is opposite in direction to the diffusion current (Fig. 15.10) starts.

Initially, diffusion current is large and drift current is small. As the diffusion process continues, the space-charge regions on either side of the junction extend, thus increasing the electric field strength and hence drift current. This process continues until the diffusion current equals the drift current. Thus a p-n junction is formed. In a p-n junction under equilibrium there is *no net* current.

The loss of electrons from the n-region and the gain of electron by the p-region causes a difference of potential across the junction of the two regions. The polarity of this potential is such as to oppose further flow of carriers so that a condition of equilibrium exists. Figure 15.11 shows the p-n junction at equilibrium and the potential across the junction. The n-material has lost electrons, and p material has acquired electrons. The n material is thus positive relative to the p material. Since this potential tends to prevent the movement of electron from the n region into the p region, it is often called a *barrier potential*.

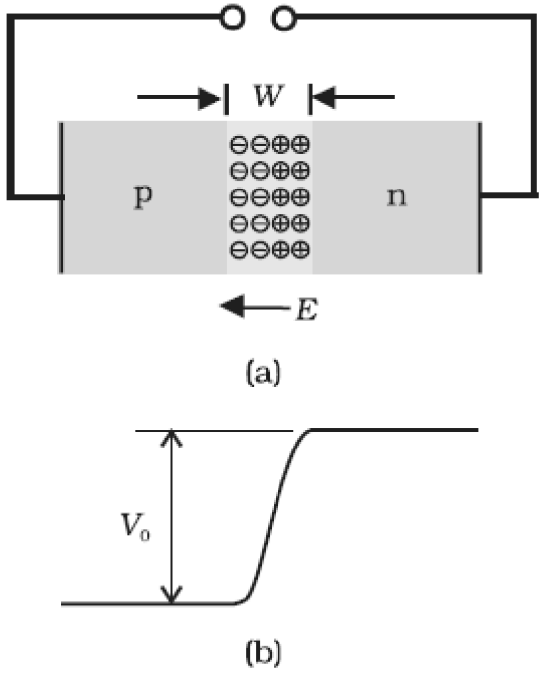


Fig. 15.11 (a) Diode under equilibrium (V = 0), (b) Barrier potential under no bias.

15.6 SEMICONDUCTOR DIODE

A semiconductor diode [Fig. 15.12(a)] is basically a p-n junction with metallic contacts provided at the ends for the application of an external voltage. It is a two terminal device. A p-n junction diode is symbolically represented as shown in Fig. 15.12(b).

The direction of arrow indicates the conventional direction of current (when the diode is under forward bias). The equilibrium barrier potential can be altered by applying an external

voltage *V* across the diode. The situation of p-n junction diode under equilibrium (without bias) is shown in Fig. 15.11(a) and (b).

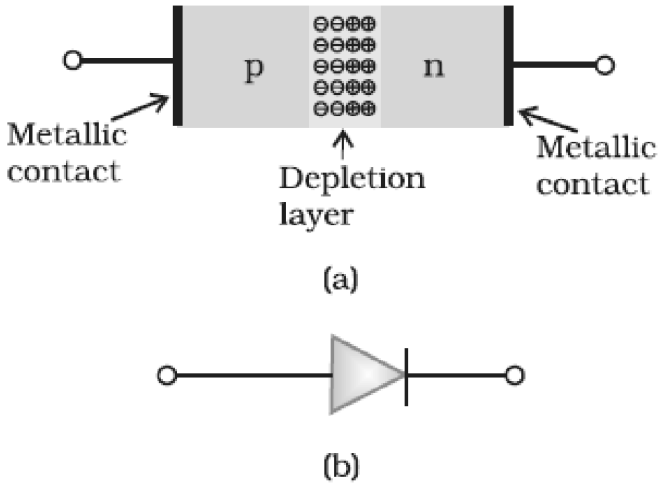


Fig. 15.12 (a) Semiconductor diode, (b) Symbol for p-n junction diode.

15.6.1 p-n junction diode under forward bias

When an external voltage *V* is applied across a semiconductor diode such that p-side is connected to the positive terminal of the battery and n-side to the negative terminal [Fig. 15.13(a)], it is said to be *forward biased.*

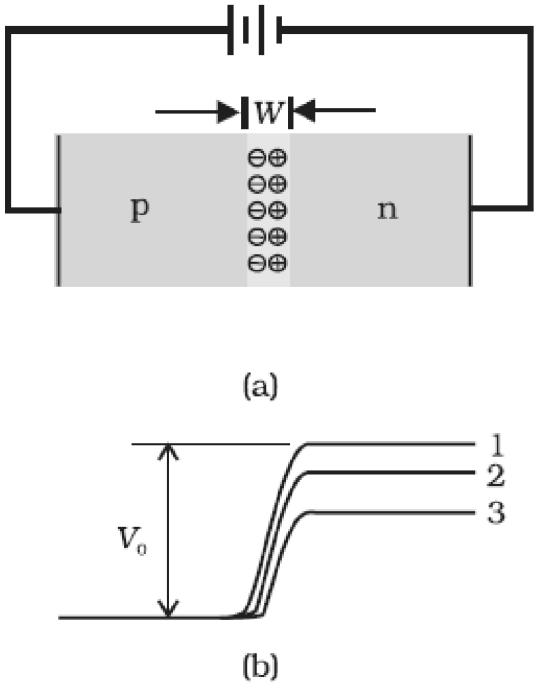


Fig. 15.13 (a) p-n junction diode under forward bias, (b) Barrier potential (1) without battery, (2) Low battery voltage, and (3) High voltage battery.

The applied voltage mostly drops across the depletion region and the voltage drop across the p-side and n-side of the junction is negligible. (This is because the resistance of the depletion region – a region where there are no charges – is very high compared to the resistance of n-side and p-side.) The direction of the applied voltage (V) is opposite to the built-in potential V_0. As a result, the depletion layer width decreases and the barrier height is reduced [Fig. 15.13(b)]. The effective barrier height under forward bias is ($V_0 - V$).

If the applied voltage is small, the barrier potential will be reduced only slightly below the equilibrium value, and only a small number of carriers in the material—those that happen to be in the uppermost energy levels—will possess enough energy to cross the junction. So the current will be small. If we increase the applied voltage significantly, the barrier height will be reduced and more number of carriers will have the required energy. Thus the current increases.

Due to the applied voltage, electrons from n-side cross the depletion region and reach p-side (where they are minority carries). Similarly, holes from p-side cross the junction and reach the n-side (where they are minority carries). This process under forward bias is known as minority carrier injection. At the junction boundary, on each side, the minority carrier concentration increases significantly compared to the locations far from the junction.

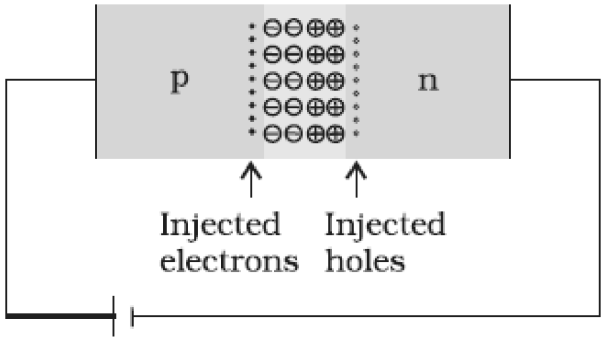


Fig. 15.14 Forward bias minority carrier injection.

Due to this concentration gradient, the injected electrons on p-side diffuse from the junction edge of p-side to the other end of p-side. Likewise, the injected holes on n-side diffuse from the junction edge of n-side to the other end of n-side (Fig. 15.14). This motion of charged carriers on either side gives rise to current. The total diode forward current is sum of hole diffusion current and conventional current due to electron diffusion. The magnitude of this current is usually in mA.

15.6.2 p-n junction diode under reverse bias

When an external voltage (V) is applied across the diode such that n-side is positive and p-side is negative, it is said to be reverse biased [Fig.15.15(a)]. The applied voltage mostly drops across the depletion region. The direction of applied voltage is same as the direction of barrier potential. As a result, the barrier height increases and the depletion region widens due to the change in the electric field. The effective barrier height under reverse bias is (V0 + V), [Fig. 15.15(b)].This suppresses the flow of electrons from n → p and holes from p → n. Thus, diffusion current, decreases enormously compared to the diode under forward bias.

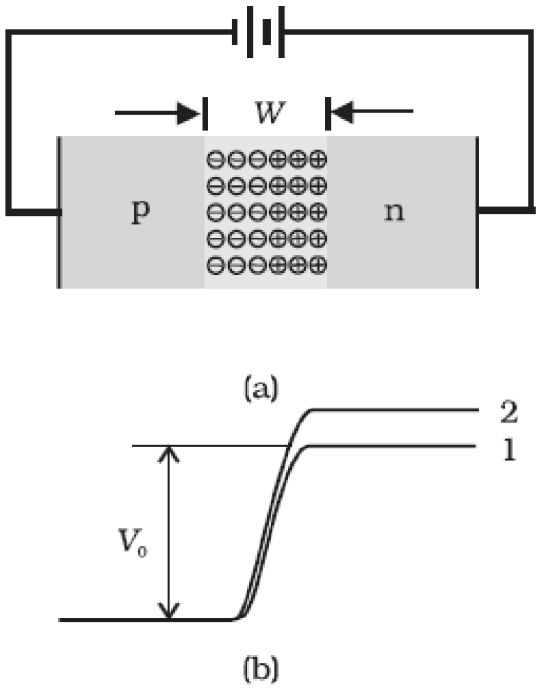


Fig. 15.15 (a) Diode under reverse bias,(b) Barrier potential under reverse bias

The electric field direction of the junction is such that if electrons on p-side or holes on n-side in their random motion come close to the junction, they will be swept to its majority zone. This drift of carriers gives rise to current. The drift current is of the order of a few μA. This is quite low because it is due to the motion of carriers from their minority side to their majority side across the junction. The drift current is also there under forward bias but it is negligible (μA) when compared with current due to injected carriers which is usually in mA.

The diode reverse current is not very much dependent on the applied voltage. Even a small voltage is sufficient to sweep the minority carriers from one side of the junction to the other side of the junction. The current is not limited by the magnitude of the applied voltage but is limited due to the concentration of the minority carrier on either side of the junction. The current under reverse bias is essentially voltage independent upto a critical reverse bias voltage, known as breakdown voltage (V_{br}). When $V = V_{br}$, the diode reverse current increases sharply. Even a slight increase in the bias voltage causes large change in the current. If the reverse current is not limited by an external circuit below the rated value (specified by the manufacturer) the p-n junction will get destroyed. Once it exceeds the rated value, the diode gets destroyed due to overheating. This can happen even for the diode under forward bias, if the forward current exceeds the rated value.

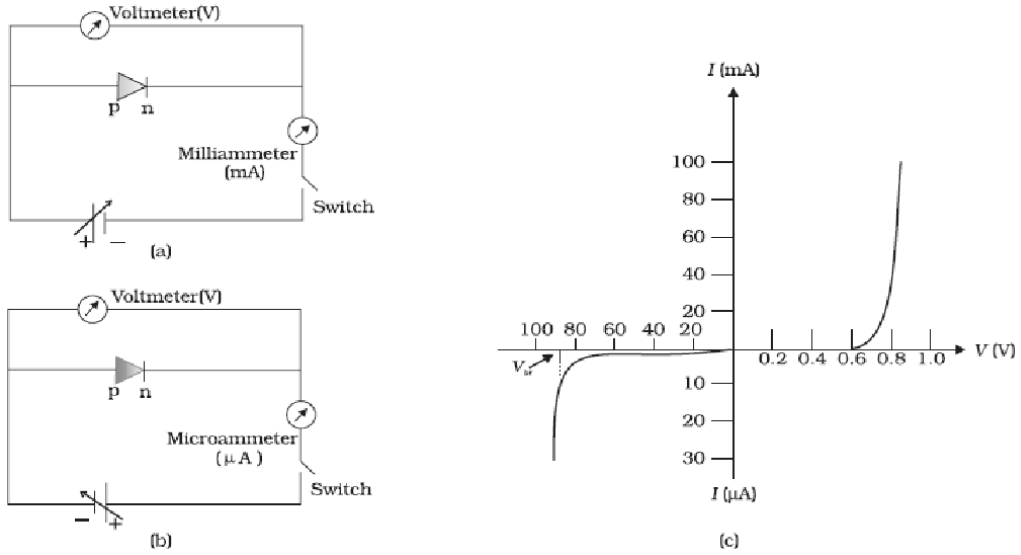


Fig. 15.16 Experimental circuit arrangement for studying *V-I* characteristics of a p-n junction diode (a) in forward bias, (b) in reverse bias, (c) Typical *V-I* characteristics of a silicon diode.

The circuit arrangement for studying the *V-I* characteristics of a diode, (i.e., the variation of current as a function of applied voltage) are shown in Fig. 15.16(a) and (b). The battery is connected to the diode through a potentiometer (or reheostat) so that the applied voltage to the diode can be changed. For different values of voltages, the value of the current is noted. A graph between *V* and *I* is obtained as in Fig. 15.16(c). Note that in forward bias measurement, we use a milliammeter since the expected current is large (as explained in the earlier section) while a micrometer is used in reverse bias to measure the current. You can see in Fig. 15.16(c) that in forward bias, the current first increases very slowly, almost negligibly, till the voltage across the diode crosses a certain value. After the characteristic voltage, the diode current increases significantly (exponentially), even for a very small increase in the diode bias voltage. This voltage is called the *threshold voltage* or cut-in voltage (~0.2V for germanium diode and ~0.7 V for silicon diode).

For the diode in reverse bias, the current is very small (~μA) and almost remains constant with change in bias. It is called *reverse saturation current*. However, for special cases, at very high reverse bias (break down voltage), the current suddenly increases. This special action of the diode is discussed later in Section 15.8. The general purpose diode are not used beyond the reverse saturation current region.

The above discussion shows that the p-n junction diode primerly allows the flow of current only in one direction (forward bias). The forward bias resistance is low as compared to the reverse bias resistance. This property is used for rectification of ac voltages as discussed in the next section. For diodes, we define a quantity called *dynamic resistance* as the ratio of small change in voltage ΔV to a small change in current ΔI:

$$r_d = \frac{\Delta V}{\Delta I} \qquad (15.6)$$

15.7 APPLICATION OF JUNCTION DIODE AS A RECTIFIER

From the *V-I* characteristic of a junction diode we see that it allows current to pass only when it is forward biased. So if an alternating voltage is applied across a diode the current flows only in that part of the cycle when the diode is forward biased. This property is used *to rectify* alternating voltages and the circuit used for this purpose is called a *rectifier*.

If an alternating voltage is applied across a diode in series with a load, a pulsating voltage will appear across the load only during the half cycles of the ac input during which the diode is forward biased. Such rectifier circuit, as shown in Fig. 15.17, is called a *half-wave rectifier*. The secondary of a transformer supplies the desired ac voltage across terminals A and B. When the voltage at A is positive, the diode is forward biased and it conducts. When A is negative, the diode is reverse-biased and it does not conduct. The reverse saturation current of a diode is negligible and can be considered equal to zero for practical purposes. (The reverse breakdown voltage of the diode must be sufficiently higher than the peak ac voltage at the secondary of the transformer to protect the diode from reverse breakdown.)

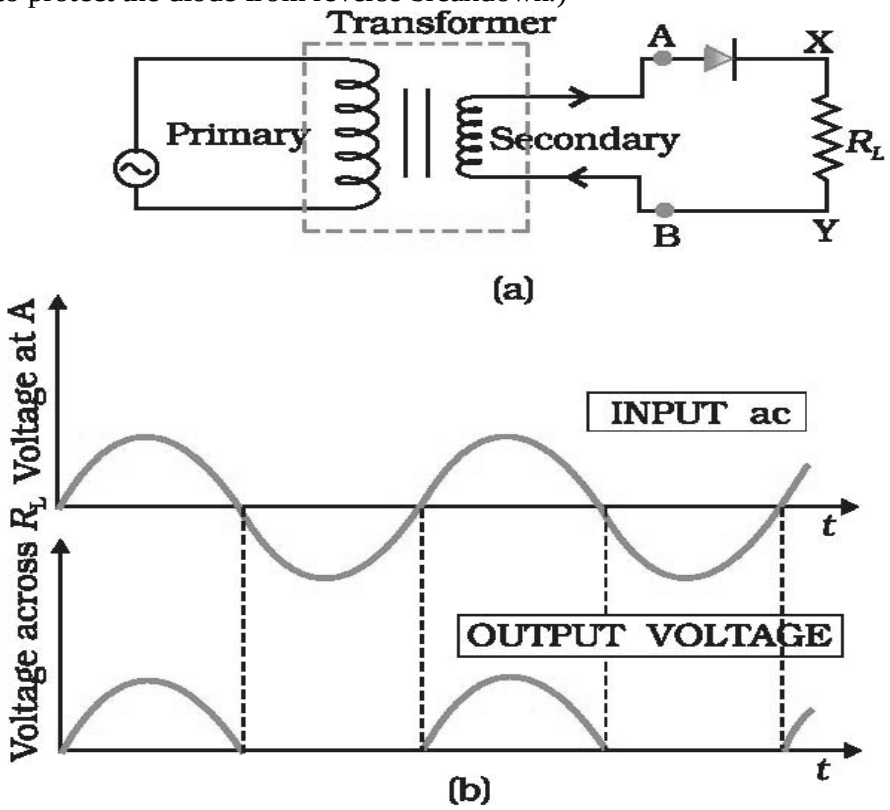


Fig. 15.17 (a) Half-wave rectifier circuit, (b) Input ac voltage and output voltage waveforms from the rectifier circuit.

Therefore, in the positive *half-cycle* of ac there is a current through the load resistor R_L and we get an output voltage, as shown in Fig. 15.17(b), whereas there is no current in the negative half-cycle. In the next positive half-cycle, again we get the output voltage. Thus, the output voltage, though still varying, is restricted to *only one direction* and is said to be *rectified*.

Since the rectified output of this circuit is only for half of the input ac wave it is called as *half-wave rectifier*.

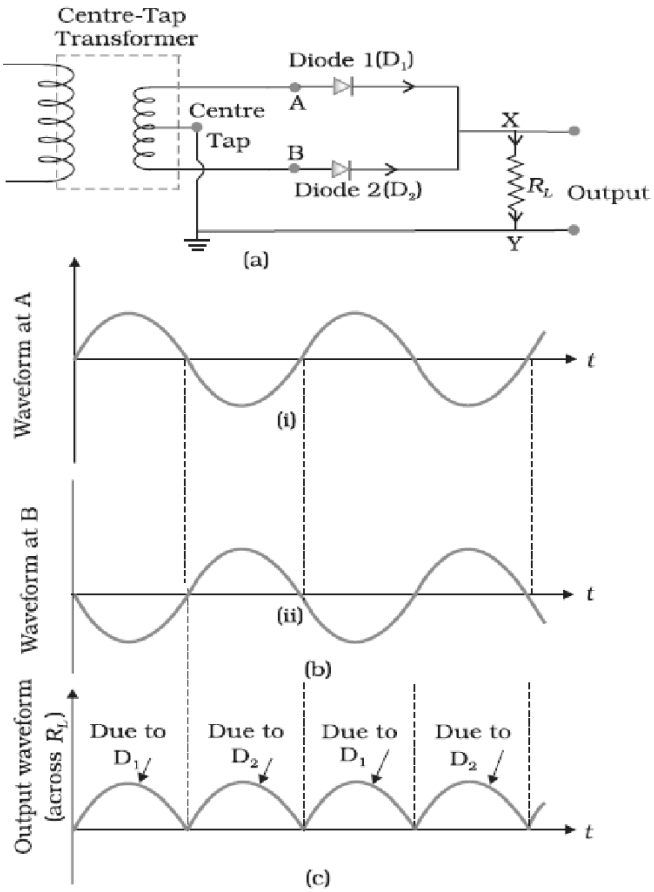


Fig. 15.18 (a) A Full-wave rectifier circuit; (b) Input wave forms given to the diode D_1 at A and to the diode D_2 at B; (c) Output waveform across the load R_L connected in the full-wave rectifier circuit.

The circuit using two diodes, shown in Fig. 15.18(a), gives output rectified voltage corresponding to both the positive as well as negative half of the ac cycle. Hence, it is known as *full-wave rectifier*. Here the p-side of the two diodes are connected to the ends of the secondary of the transformer. The n-side of the diodes are connected together and the output is taken between this common point of diodes and the midpoint of the secondary of the transformer. So for a full-wave rectifier the secondary of the transformer is provided with a centre tapping and so it is called *centre-tap transformer*. As can be seen from Fig.15.18(c) the voltage rectified by each diode is only half the total secondary voltage. Each diode rectifies only for half the cycle, but the two do so for alternate cycles. Thus, the output between their common terminals and the centre-tap of the transformer becomes a full-wave rectifier output. (Note that there is another circuit of full wave rectifier which does not need a centre tap transformer but needs four diodes.) Suppose the input voltage to A with respect to the centre tap at any instant is positive. It is clear that, at

that instant, voltage at B being out of phase will be negative as shown in Fig. 15.18(b). So, diode D_1 gets forward biased and conducts (while D_2 being reverse biased is not conducting).

Hence, during this positive half cycle we get an output current (and a output voltage across the load resistor R_L) as shown in Fig.15.18(c). In the course of the ac cycle when the voltage at A becomes negative with respect to centre tap, the voltage at B would be positive. In this part of the cycle diode D_1 would not conduct but diode D_2 would, giving an output current and output voltage (across R_L) during the negative half cycle of the input ac. Thus, we get output voltage during both the positive as well as the negative half of the cycle. Obviously, this is a more efficient circuit for getting rectified voltage or current than the half-wave rectifier.

The rectified voltage is in the form of pulses of the shape of half sinusoids. Though it is unidirectional it does not have a steady value. To get steady dc output from the pulsating voltage normally a capacitor is connected across the output terminals (parallel to the load R_L). One can also use an inductor in series with R_L for the same purpose. Since these additional circuits appear to *filter* out the *ac ripple* and give a *pure dc* voltage, so they are called filters.

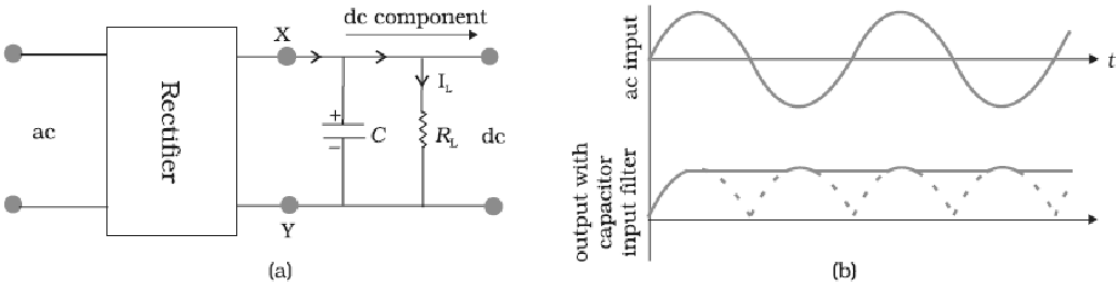


Fig. 15.19 (a) A full-wave rectifier with capacitor filter, (b) Input and output voltage of rectifier in (a).

Now we shall discuss the role of capacitor in filtering. When the voltage across the capacitor is rising, it gets charged. If there is no external load, it remains charged to the peak voltage of the rectified output. When there is a load, it gets discharged through the load and the voltage across it begins to fall. In the next half-cycle of rectified output it again gets charged to the peak value (Fig. 15.19). The rate of fall of the voltage across the capacitor depends upon the inverse product of capacitor C and the effective resistance R_L used in the circuit and is called the *time constant*. To make the time constant large value of C should be large. So capacitor input filters use large capacitors. The *output voltage* obtained by using capacitor input filter is nearer to the *peak voltage* of the rectified voltage. This type of filter is most widely used in power supplies.

15.8 SPECIAL PURPOSE p-n JUNCTION DIODES

In the section, we shall discuss some devices which are basically junction diodes but are developed for different applications.

15.8.1 Zener diode

It is a special purpose semiconductor diode, named after its inventor C. Zener. It is designed to operate under reverse bias in the breakdown region and used as a voltage regulator. The symbol for Zener diode is shown in Fig. 15.20(a). Zener diode is fabricated by heavily doping both p-, and n- sides of the junction. Due to this, depletion region formed is very thin ($<10^{-6}$ m) and the

electric field of the junction is extremely high (~5 × 10^6 V/m) even for a small reverse bias voltage of about 5V. The I-V characteristics of a Zener diode is shown in Fig. 15.20(b).

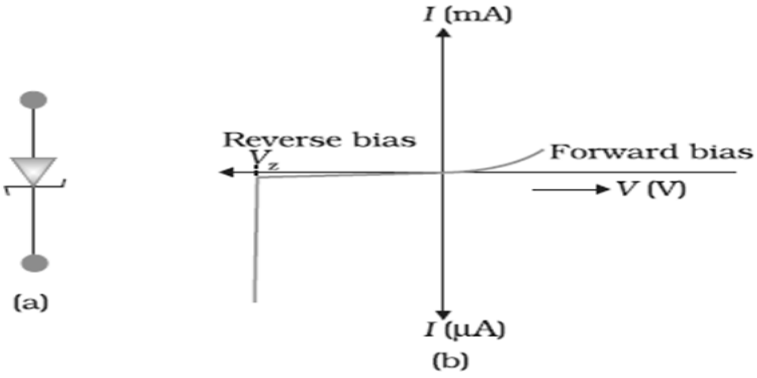


Fig. 15.20 Zener diode, (a) symbol, (b) *I-V* characteristics.

It is seen that when the applied reverse bias voltage (*V*) reaches the breakdown voltage (V_z) of the Zener diode, here is a large change in the current. Note that after the breakdown voltage V_z, a large change in the current can be produced by almost insignificant change in the reverse bias voltage. In other words, Zener voltage remains constant, even though current through the Zener diode varies over a wide range. This property of the Zener diode is used for regulating supply voltages so that they are constant.

Let us understand how reverse current suddenly increases at the breakdown voltage. We know that reverse current is due to the flow of electrons (minority carriers) from p →n and holes from n →p. As the reverse bias voltage is increased, the electric field at the junction becomes significant. When the reverse bias voltage V = V_z, then the electric field strength is high enough to pull valence electrons from the host atoms on the p-side which are accelerated to n-side. These electrons account for high current observed at the breakdown. The emission of electrons from the host atoms due to the high electric field is known as internal field emission or field ionisation. The electric field required for field ionisation is of the order of 10^6 V/m.

Zener diode as a voltage regulator

We know that when the ac input voltage of a rectifier fluctuates, its rectified output also fluctuates. To get a constant dc voltage from the dc unregulated output of a rectifier, we use a Zener diode. The circuit diagram of a voltage regulator using a Zener diode is shown in Fig. 15.21.

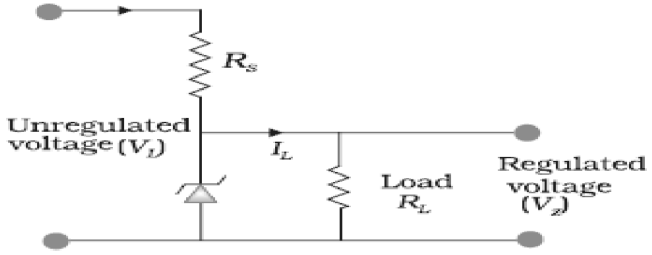


Fig. 15.21 Zener diode as DC voltage regulator (to be corrected).

The unregulated dc voltage (filtered output of a rectifier) is connected to the Zener diode through a series resistance R_s such that the Zener diode is reverse biased. If the input voltage increases, the current through R_s and Zener diode also increases. This increases the voltage drop across R_s without any change in the voltage across the Zener diode. This is because in the breakdown region, Zener voltage remains constant even though the current through the Zener diode changes. Similarly, if the input voltage decreases, the current through *Rs* and Zener diode also decreases. The voltage drop across R_s decreases without any change in the voltage across the Zener diode. Thus any increase/ decrease in the input voltage results in, increase/ decrease of the voltage drop across *Rs* without any change in voltage across the Zener diode. Thus the Zener diode acts as a voltage regulator. We have to select the Zener diode according to the required output voltage and accordingly the series resistance R_s.

15.8.2 Optoelectronic junction devices

We have seen so far, how a semiconductor diode behaves under applied electrical inputs. In this section, we learn about semiconductor diodes in which carriers are generated by photons (photo-excitation). All these devices are called *optoelectronic devices*. We shall study the functioning of the following optoelectronic devices:

(i) *Photodiodes* used for detecting optical signal (photodetectors).

(ii) *Light emitting diodes* (LED) which convert electrical energy into light.

(iii) *Photovoltaic devices* which convert optical radiation into electricity (*solar cells*).

(i) *Photodiode*

A Photodiode is again a special purpose p-n junction diode fabricated with a transparent window to allow light to fall on the diode. It is operated under reverse bias. When the photodiode is illuminated with light (photons) with energy (hν) greater than the energy gap (E_g) of the semiconductor, then electron-hole pairs are generated due to the absorption of photons. The diode is fabricated such that the generation of e-h pairs takes place in or near the depletion region of the diode. Due to electric field of the junction, electrons and holes are separated before they recombine. The direction of the electric field is such that electrons reach n-side and holes reach p-side. Electrons are collected on n-side and holes are collected on p-side giving rise to an emf. When an external load is connected, current flows. The magnitude of the photocurrent depends on the intensity of incident light (photocurrent is proportional to incident light intensity).

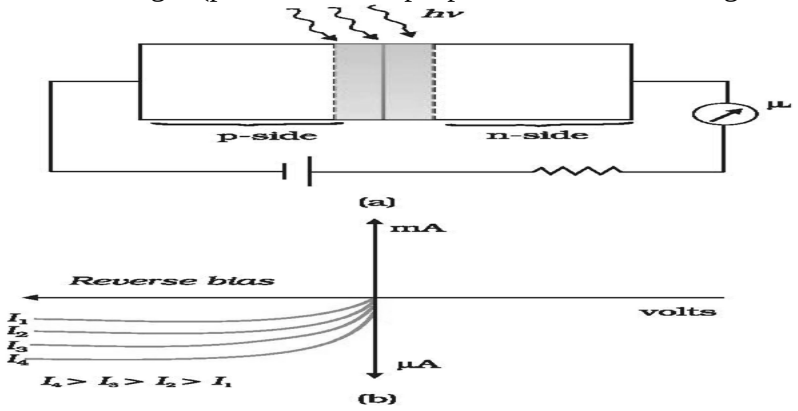


Fig. 15.22 (a) An illuminated photodiode under reverse bias, (b) *I-V* characteristics of a photodiode for different illumination intensity $I_4 > I_3 > I_2 > I_1$.

It is easier to observe the change in the current with change in the light intensity, if a reverse bias is applied. Thus photodiode can be used as a photodetector to detect optical signals. The circuit diagram used for the measurement of I-V characteristics of a photodiode is shown in Fig. 15.22(a) and typical I-V characteristics in Fig. 15.22(b).

(ii) ***Light emitting diode***

It is a heavily doped p-n junction which under forward bias emits spontaneous radiation. The diode is encapsulated with a transparent cover so that emitted light can come out.

When the diode is forward biased, electrons are sent from n → p (where they are minority carriers) and holes are sent from p → n (where they are minority carriers). At the junction boundary the concentration of minority carriers increases compared to the equilibrium concentration (i.e., when there is no bias). Thus at the junction boundary on either side of the junction, excess minority carriers are there which recombine with majority carriers near the junction. On recombination, the energy is released in the form of photons. Photons with energy equal to or slightly less than the band gap are emitted. When the forward current of the diode is small, the intensity of light emitted is small. As the forward current increases, intensity of light increases and reaches a maximum. Further increase in the forward current results in decrease of light intensity. LEDs are biased such that the light emitting efficiency is maximum.

The *V-I* characteristics of a LED is similar to that of a Si junction diode. But the threshold voltages are much higher and slightly different for each colour. The reverse breakdown voltages of LEDs are very low, typically around 5V. So care should be taken that high reverse voltages do not appear across them.

LEDs that can emit red, yellow, orange, green and blue light are commercially available. The semiconductor used for fabrication of visible LEDs must at least have a band gap of 1.8 eV (spectral range of visible light is from about 0.4 μm to 0.7 μm, i.e., from about 3 eV to 1.8 eV). The compound semiconductor Gallium Arsenide – Phosphide ($GaAs_{1-x}P_x$) is used for making LEDs of different colours. $GaAs_{0.6}P_{0.4}$ (E_g ~ 1.9 eV) is used for red LED. GaAs (E_g ~ 1.4 eV) is used for making infrared LED. These LEDs find extensive use in remote controls, burglar alarm systems, optical communication, etc. Extensive research is being done for developing white LEDs which can replace incandescent lamps.

LEDs have the following advantages over conventional incandescent low power lamps:

(i) Low operational voltage and less power.

(ii) Fast action and no warm-up time required

(iii) The bandwidth of emitted light is 100 Å to 500 Å or in other words it is nearly (but not exactly) monochromatic.

(iv) Long life and ruggedness.

(v) Fast on-off switching capability.

(iii) ***Solar cell***

A solar cell is basically a p-n junction which generates emf when solar radiation falls on the p-n junction. It works on the same principle (photovoltaic effect) as the photodiode, except that no external bias is applied and the junction area is kept much larger for solar radiation to be incident because we are interested in more power. A simple p-n junction solar cell is shown in Fig. 15.23.

A p-Si wafer of about 300 μm is taken over which a thin layer (~0.3 μm) of n-Si is grown on one-side by diffusion process. The other side of p-Si is coated with a metal (back contact). On the top of n-Si layer, metal finger electrode (or metallic grid) is deposited. This acts as a front contact. The metallic grid occupies only a very small fraction of the cell area (<15%) so that light can be incident on the cell from the top.

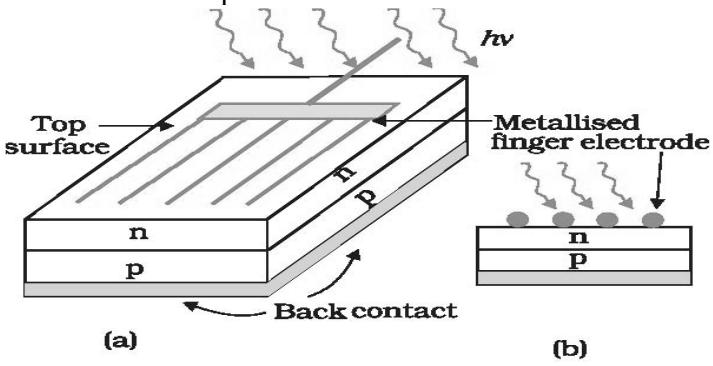


Fig. 15.23 (a) Typical p-n junction solar cell; (b) Cross-sectional view.

The generation of emf by a solar cell, when light falls on, it is due to the following three basic processes: generation, separation and collection-

(i) generation of e-h pairs due to light (with $h\nu > E_g$) close to the junction;

(ii) separation of electrons and holes due to electric field of the depletion region. Electrons are swept to n-side and holes to p-side;

(iii) the electrons reaching the n-side are collected by the front contact and holes reaching p-side are collected by the back contact. Thus p-side becomes positive and n-side becomes negative giving rise to *photovoltage*.

When an external load is connected as shown in the Fig. 15.24(a) a photocurrent I_L flows through the load. A typical *I-V* characteristics of a solar cell is shown in the Fig. 15.24(b).

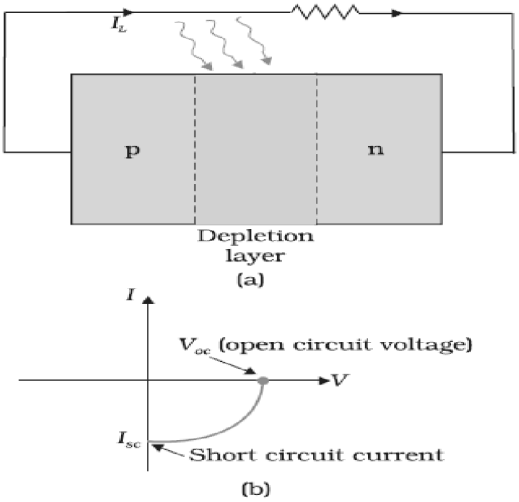


Fig. 15.24 (a) A typical illuminated p-n junction solar cell; (b) *I-V* characteristics of a solar cell.

Note that the $I - V$ characteristics of solar cell is drawn in the fourth quadrant of the coordinate axes. This is because a solar cell does not draw current but supplies the same to the load.

Semiconductors with band gap close to 1.5 eV are ideal materials for solar cell fabrication. Solar cells are made with semiconductors like Si (E_g = 1.1 eV), GaAs (E_g = 1.43 eV), CdTe (E_g = 1.45 eV), $CuInSe_2$ (E_g = 1.04 eV), etc. The important criteria for the selection of a material for solar cell fabrication are (i) band gap (~1.0 to 1.8 eV), (ii) high optical absorption (~10^4 cm^{-1}), (iii) electrical conductivity, (iv) availability of the raw material, and (v) cost. Note that sunlight is not always required for a solar cell. *Any* light with photon energies greater than the bandgap will do. Solar cells are used to power electronic devices in satellites and space vehicles and also as power supply to some calculators. Production of low-cost photovoltaic cells for large-scale solar energy is a topic for research.

15.9 DIGITAL ELECTRONICS AND LOGIC GATES

In electronics circuits like amplifiers, oscillators, introduced to you in earlier sections, the signal (current or voltage) has been in the form of continuous, time-varying voltage or current. Such signals are called continuous or *analogue signals*. A typical analogue signal is shown in Figure. 15.25(a). Fig. 15.25(b) shows a *pulse waveform* in which only discrete values of voltages are possible. It is convenient to use binary numbers to represent such signals. A binary number has only two digits '0' (say, 0V) and '1' (say, 5V). In digital electronics we use only these two levels of voltage as shown in Fig. 15.25(b). Such signals are called *Digital Signals*. In digital circuits only two values (represented by 0 or 1) of the input and output voltage are permissible.

This section is intended to provide the first step in our understanding of digital electronics. We shall restrict our study to some basic building blocks of digital electronics (called *Logic Gates*) which process the digital signals in a specific manner. Logic gates are used in calculators, digital watches, computers, robots, industrial control systems, and in telecommunications.

A light switch in your house can be used as an example of a digital circuit. The light is either ON or OFF depending on the switch position. When the light is ON, the output value is '1'. When the light is OFF the output value is '0'. The inputs are the position of the light switch. The switch is placed either in the ON or OFF position to activate the light.

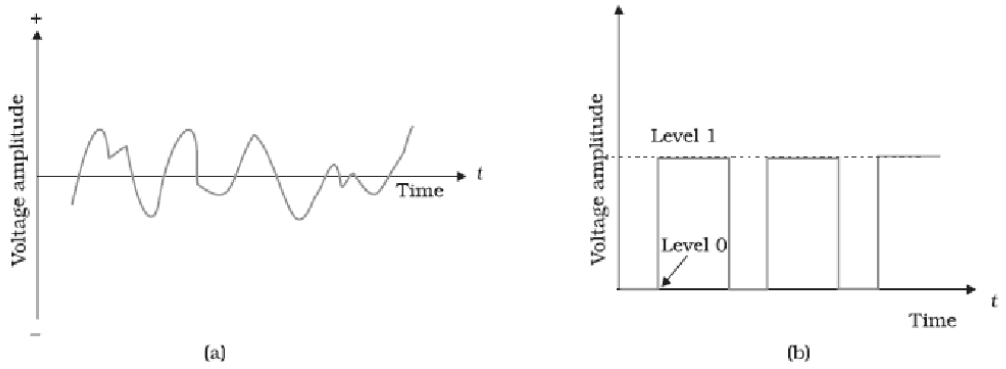
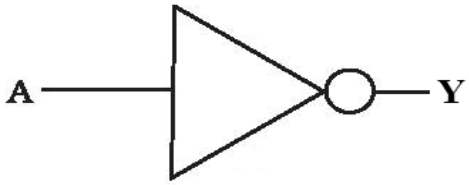


Fig. 15.25 (a) Analogue signal, (b) Digital signal.

15.9.1 Logic gates



Input	Output
A	Y
0	1
1	0

(a) (b)

Fig.15.26 (a) Logic symbol, (b) Truth table of NOT gate.

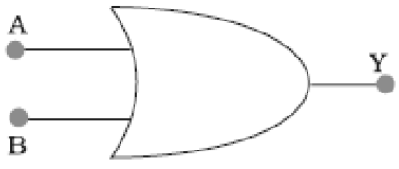
A gate is a digital circuit that follows curtain *logical* relationship between the input and output voltages. Therefore, they are generally known as *logic gates* — gates because they control the flow of information. The five common logic gates used are NOT, AND, OR, NAND, NOR. Each logic gate is indicated by a symbol and its function is defined by a *truth table* that shows all the possible input logic level combinations with their respective output logic levels. Truth tables help understand the behaviour of logic gates. These logic gates can be realised using semiconductor devices.

(i) NOT gate

This is the most basic gate, with one input and one output. It produces a '1' output if the input is '0' and vice-versa. That is, it produces an inverted version of the input at its output. This is why it is also known as an *inverter*. The commonly used symbol together with the truth table for this gate is given in Fig. 15.26.

(ii) OR Gate

An *OR* gate has two or more inputs with one output. The logic symbol and truth table are shown in Fig. 15.27. The output Y is 1 when either input A *or* input B *or* both are 1s, that is, if any of the input is high, the output is high.



Input		Output
A	B	Y
0	0	0
0	1	1
1	0	1
1	1	1

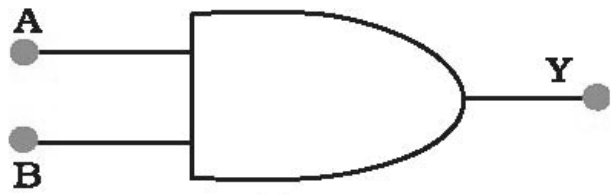
(a) (b)

Fig. 15.27 (a) Logic symbol (b) Truth table of OR gate.

Apart from carrying out the above mathematical logic operation, this *gate* can be used for modifying the pulse waveform as explained in the following example.

(iii) AND Gate

An *AND* gate has two or more inputs and one output. The output Y of AND gate is 1 only when input A *and* input B are both 1. The logic symbol and truth table for this gate are given in Fig. 15.28



(a)

Input		Output
A	B	Y
0	0	0
0	1	0
1	0	0
1	1	1

(b)

Fig. 15.28 (a) Logic symbol, (b) Truth table of AND gate.

(iv) NAND Gate

This is an AND gate followed by a NOT gate. If inputs A *and* B are both '1', the output Y is *not* '1'. The gate gets its name from this NOT AND behaviour. Fig. 15.29 shows the symbol and truth table of NAND gate.

NAND gates are also called *Universal Gates* since by using these gates you can realise other basic gates like OR, AND and NOT

(a)

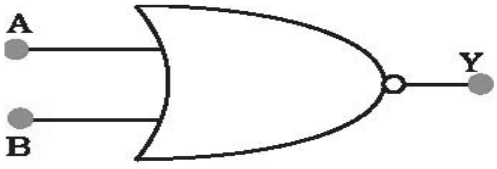
Input		Output
A	B	Y
0	0	1
0	1	1
1	0	1
1	1	0

(b)

Fig. 15.29 (a) Logic symbol, (b) Truth table of NAND gate.

(*v) NOR Gate*

It has two or more inputs and one output. A NOT- operation applied *after* OR gate gives a NOT-OR gate (or simply NOR gate). Its output Y is '1' only when both inputs A and B are '0', i.e., neither one input *nor* the other is '1'. The symbol and truth table for NOR gate is given in Fig. 15.30.



(a)

Input		Output
A	B	Y
0	0	1
0	1	0
1	0	0
1	1	0

(b)

Fig. 15.30 (a) Logic symbol, (b) Truth table of NOR gate.

NOR gates are considered as *universal* gates because you can obtain all the gates like AND, OR, NOT by using only NOR gates.

FASTER AND SMALLER: THE FUTURE OF COMPUTER TECHNOLOGY

The Integrated Chip (IC) is at the heart of all computer systems. In fact ICs are found in almost all electrical devices like cars, televisions, CD players, cell phones etc. The miniaturisation that made the modern personal computer possible could never have happened without the IC. ICs are electronic devices that contain many transistors, resistors, capacitors, connecting wires – all in one package. You must have heard of the microprocessor. The microprocessor is an IC that processes all information in a computer, like keeping track of what keys are pressed, running programmes, games etc. The IC was first invented by Jack Kilky at Texas Instruments in 1958 and he was awarded Nobel Prize for this in 2000. ICs are produced on a piece of semiconductor crystal (or chip) by a process called photolithography. Thus, the entire Information Technology (IT) industry hinges on semiconductors. Over the years, the complexity of ICs has increased while the size of its features continued to shrink. In the past five decades, a dramatic miniaturisation in computer technology has made modern day computers faster and smaller. In the 1970s, Gordon Moore, co-founder of INTEL, pointed out that the memory capacity of a chip (IC) approximately doubled every one and a half years. This is popularly known as Moore's law. The number of transistors per chip has risen exponentially and each year computers are becoming more powerful, yet cheaper than the year before. It is intimated from current trends that the computers available in 2020 will operate at 40 GHz (40,000 MHz) and would be much smaller, more efficient and less expensive than present day computers. The explosive growth in the semiconductor industry and computer technology is best expressed by a famous quote from Gordon Moore: "If the auto industry advanced as rapidly as the semiconductor industry, a Rolls Royce would get half a million miles per gallon, and it would be cheaper to throw it away than to park it".

SUMMARY

1. Semiconductors are the basic materials used in the present solid state electronic devices like diode, transistor, ICs, etc.
2. Lattice structure and the atomic structure of constituent elements decide whether a particular material will be insulator, metal or semiconductor.
3. Metals have low resistivity (10^{-2} to 10^{-8} Ωm), insulators have very high resistivity ($>10^{8}$ Ωm^{-1}), while semiconductors have intermediate values of resistivity.
4. Semiconductors are elemental (Si, Ge) as well as compound (GaAs,CdS, etc.).
5. Pure semiconductors are called 'intrinsic semiconductors'. The presence of charge carriers (electrons and holes) is an 'intrinsic' property of the material and these are obtained as a result of thermal excitation. The number of electrons (n_e) is equal to the number of holes (n_h) in intrinsic conductors. Holes are essentially electron vacancies with an effective positive charge.
6. The number of charge carriers can be changed by 'doping' of a suitable impurity in pure semiconductors. Such semiconductors are known as extrinsic semiconductors. These are of two types (n-type and p-type).
7. In n-type semiconductors, $n_e >> n_h$ while in p-type semiconductors $n_h >> n_e$.
8. n-type semiconducting Si or Ge is obtained by doping with pentavalent atoms (donors) like As, Sb, P, etc., while p-type Si or Ge can be obtained by doping with trivalent atom (acceptors) like B, Al, In etc.
9. $n_e n_h = n_i^2$ in all cases. Further, the material possesses an *overall charge neutrality*.
10. There are two distinct band of energies (called valence band and conduction band) in which the electrons in a material lie. Valence band energies are low as compared to conduction band energies. All energy levels in the valence band are filled while energy levels in the conduction band may be fully empty or partially filled. The electrons in the conduction band are free to move in a solid and are responsible for the conductivity. The extent of conductivity depends upon the energy gap (E_g) between the top of valence band (E_V) and the bottom of the conduction band E_C. The electrons from valence band can be excited by heat, light or electrical energy to the conduction band and thus, produce a change in the current flowing in a semiconductor.
11. For insulators $E_g > 3$ eV, for semiconductors E_g is 0.2 eV to 3 eV, while for metals $E_g \approx 0$.
12. p-n junction is the 'key' to all semiconductor devices. When such a junction is made, a 'depletion layer' is formed consisting of immobile ion-cores devoid of their electrons or holes. This is responsible for a junction potential barrier.
13. By changing the external applied voltage, junction barriers can be changed. In forward bias (n-side is connected to negative terminal of the battery and p-side is connected to the positive), the barrier is decreased while the barrier increases in reverse bias. Hence, forward bias current is more (mA) while it is very small (μA) in a p-n junction diode.
14. Diodes can be used for rectifying an ac voltage (restricting the ac voltage to one direction). With the help of a capacitor or a suitable filter, a dc voltage can be obtained.
15. There are some special purpose diodes.

16. Zener diode is one such special purpose diode. In reverse bias, after a certain voltage, the current suddenly increases (breakdown voltage) in a Zener diode. This property has been used to obtain *voltage regulation*.
17. p-n junctions have also been used to obtain many photonic or optoelectronic devices where one of the participating entity is 'photon': (a) Photodiodes in which photon excitation results in a change of reverse saturation current which helps us to measure light intensity; (b) Solar cells which convert photon energy into electricity; (c) Light Emitting Diode and Diode Laser in which electron excitation by a bias voltage results in the generation of light.
18. There are some special circuits which handle the digital data consisting of 0 and 1 levels. This forms the subject of Digital Electronics.
19. The important digital circuits performing special logic operations are called logic gates. These are: OR, AND, NOT, NAND, and NOR gates.

POINTS TO PONDER

1. The energy bands (E_C or E_V) in the semiconductors are space delocalized which means that these are not located in any specific place inside the solid. The energies are the overall averages. When you see a picture in which E_C or E_V are drawn as straight lines, then they should be respectively taken simply as the *bottom* of conduction band energy levels and *top* of valence band energy levels.
2. In elemental semiconductors (Si or Ge), the n-type or p-type semiconductors are obtained by introducing 'dopants' as defects. In compound semiconductors, the change in relative stoichiometric ratio can also change the type of semiconductor. For example, in ideal GaAs the ratio of Ga:As is 1:1 but in Ga-rich or As-rich GaAs it could respectively be $Ga_{1.1}$ $As_{0.9}$ or $Ga_{0.9}$ $As_{1.1}$. In general, the presence of defects controls the properties of semiconductors in many ways.
3. In modern day circuit, many logical gates or circuits are integrated in one single 'chip'. These are known as Integrated circuits (IC).

VERY SHORT ANSWER QUESTIONS

1. What is an n-type semiconductor? What are the majority and minority charge carriers in it?
2. What are intrinsic and extrinsic semi-conductors?
3. What is a p-type semiconductor? What are the majority and minority charge carriers in it?
4. What is a p-n junction diode? Define depletion layer.
5. How is a battery connected to a junction diode in (i) forward and (ii) reverse bias?
6. What is the maximum percentage of rectification in half wave and full wave rectifiers?
7. What is Zener voltage (V_z) and how will a Zener diode be connected in circuits generally?
8. Write the expressions for the efficiency of a full wave rectifier and a half wave rectifier.
9. What happens to the width of the depletion layer in a p-n junction diode when it is (i) forward-bioased and (ii) reverse-biased?
10. Draw the circuit symbols for p-n-p and n-p-n transistors.
11. Define amplifier and amplification factor.
12. In which bias can a Zener diode be used as voltage regulator?
13. Which gates are called universal gates?
14. Write the truth table of NAND gate. How does it differ from AND gate?

SHORT ANSWER QUESTIONS

1. What are n-type and p-type semi-conductors? How is a semi-conductor junction formed?
2. Draw and explain the current-voltage (I-V) characteristic curves of a junction diode in forward and reverse bias.
3. Describe how a semiconductor diode is used as a half wave rectifier.
4. What is rectification? Explain the working of a full wave rectifier.
5. Distinguish between half-wave and full wave rectifiers.
6. Explain the different transistor configurations with diagrams.
7. Define NAND and NOR gates. Give their truth tables.

LONG ANSWER QUESTIONS

1. What is Zener diode? Explain how it is used as a voltage regulator.
2. Describe a transistor and explain its working.

CHAPTER 16

COMMUNICATION SYSTEMS

16.1 INTRODUCTION

Communication is the act of transmission of information. Every living creature in the world experiences the need to impart or receive information almost continuously with others in the surrounding world. For communication to be successful, it is essential that the sender and the receiver understand a common *language*. Man has constantly made endeavors to improve the quality of communication with other human beings. Languages and methods used in communication have kept evolving from prehistoric to modern times, to meet the growing demands in terms of speed and complexity of information. It would be worthwhile to look at the major milestones in events that promoted developments in communications, as presented in Table 16.1.

Modern communication has its roots in the 19th and 20th century in the work of scientists like J.C. Bose, F.B. Morse, G. Marconi and Alexander Graham Bell. The pace of development seems to have increased dramatically after the first half of the 20th century. We can hope to see many more accomplishments in the coming decades. The aim of this chapter is to introduce the concepts of communication, namely the mode of communication, the need for modulation, production and deduction of amplitude modulation.

16.2 ELEMENTS OF A COMMUNICATION SYSTEM

Communication pervades all stages of life of all living creatures. Irrespective of its nature, every communication system has three essential elements-transmitter, medium/ channel and receiver. The block diagram shown in Fig. 16.1 depicts the general form of a communication system.

TABLE 16.1 SOME MAJOR MILESTONES IN THE HISTORY OF COMMUNICATION

Year	Event	Remarks
Around 1565 A.D.	The reporting of the delivery of a child by queen using drum beats from a distant place to King Akbar.	It is believed that minister Birbal experimented with the arrangement to decide the number of drummers posted between the place where the queen stayed and the place where the king stayed.
1835	Invention of telegraph by Samuel F.B. Morse and Sir Charles Wheatstone	It resulted in tremendous growth of messages through post offices and reduced physical travel of messengers considerably.
1876	Telephone invented by Alexander Graham Bell and Antonio Meucci	Perhaps the most widely used means of communication in the history of mankind.
1895	Jagadis Chandra Bose and Guglielmo Marconi demonstrated wireless telegraphy.	It meant a giant leap – from an era of communication using wires to communicating without using wires. (wireless)
1936	Television broadcast (John Logi Baird)	First television broadcast by BBC

1955	First radio FAX transmitted across continent. (Alexander Bain)	The idea of FAX transmission was patented by Alexander Bain in 1843.
1968	ARPANET- the first internet came into existence (J.C.R. Licklider)	ARPANET was a project undertaken by the U.S. defence department. It allowed file transfer from one computer to another connected to the network.
1975	Fiber optics developed at Bell Laboratories	Fiber optical systems are superior and more economical compared to traditional communication systems.
1989-91	Tim Berners-Lee invented the **World Wide Web**.	WWW may be regarded as the mammoth encyclopedia of knowledge accessible to everyone round the clock throughout the year.

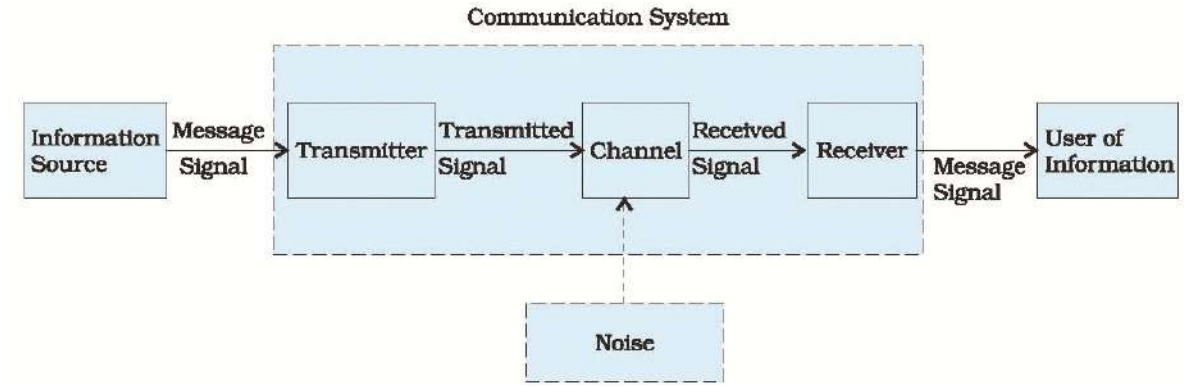


FIGURE 16.1 Block diagram of a generalised communication system.

In a communication system, the transmitter is located at one place, the receiver is located at some other place (far or near) separate from the transmitter and the channel is the physical medium that connects them. Depending upon the type of communication system, a channel may be in the form of wires or cables connecting the transmitter and the receiver or it may be wireless. The purpose of the transmitter is to convert the message signal produced by the source of information into a form suitable for transmission through the channel. If the output of the information source is a non-electrical signal like a voice signal, a transducer converts it to electrical form before giving it as an input to the transmitter. When a transmitted signal propagates along the channel it may get distorted due to channel imperfection. Moreover, noise adds to the transmitted signal and the receiver receives a corrupted version of the transmitted signal. The receiver has the task of operating on the received signal. It reconstructs a recognisable form of the original message signal for delivering it to the user of information.

There are two basic modes of communication**:** *point-to-point* and *broadcast*.

In point-to-point communication mode, communication takes place over a link between a single transmitter and a receiver. Telephony is an example of such a mode of communication. In contrast, in the broadcast mode, there are a large number of receivers corresponding to a single transmitter. Radio and television are examples of broadcast mode of communication.

16.3 BASIC TERMINOLOGY USED IN ELECTRONIC COMMUNICATION SYSTEMS

By now, we have become familiar with some terms like information source, transmitter, receiver, channel, noise, etc. It would be easy to understand the principles underlying any communication, if we get ourselves acquainted with the following basic terminology.

JAGADIS CHANDRA BOSE (1858 – 1937)

Jagadis Chandra Bose (1858–1937) He developed an apparatus for generating ultra-short electro-magnetic waves and studied their quasi-optical properties. He was said to be the first to employ a semiconductor like galena as a self- recovering detector of electromagnetic waves. Bose published three papers in the British magazine, 'The Electrician' of 27 Dec. 1895. His invention was published in the 'Proceedings of The Royal Society' on 27 April 1899 over two years before Marconi's first wireless communication on 13 December 1901. Bose also invented highly sensitive instruments for the detection of minute responses by living organisms to external stimuli and established parallelism between plant and animal tissues.

(i) ***Transducer:*** Any device that converts one form of energy into another can be termed as a transducer. In electronic communication systems, we usually come across devices that have either their inputs or outputs in the electrical form. An electrical transducer may be defined as a device that converts some physical variable (pressure, displacement, force, temperature, etc) into corresponding variations in the electrical signal at its output.

(ii) ***Signal:*** Information converted in electrical form and suitable for transmission is called a signal. Signals can be either *analog or digital*. Analog signals are continuous variations of voltage or current. *They are essentially single-valued functions of time*. Sine wave is a fundamental analog signal. All other analog signals can be fully understood in terms of their sine wave components. Sound and picture signals in TV are analog in nature. Digital signals are those which can take only discrete stepwise values. Binary system that is extensively used in digital electronics employs just two levels of a signal. '0' corresponds to a low level and '1' corresponds to a high level of voltage/ current. There are several coding schemes useful for digital communication. They employ suitable combinations of number systems such as the binary coded decimal (BCD)*****. American Standard Code for Information Interchange (ASCII)****** is a universally popular digital code to represent numbers, letters and certain characters.

(iii) ***Noise:*** Noise refers to the unwanted signals that tend to disturb the transmission and processing of message signals in a communication system. The source generating the noise may be located inside or outside the system.

(iv) ***Transmitter:*** A transmitter processes the incoming message signal so as to make it suitable for transmission through a channel and subsequent reception.

(v) ***Receiver:*** A receiver extracts the desired message signals from the received signals at the channel output.

(vi) ***Attenuation:*** The loss of strength of a signal while propagating through a medium is known as attenuation.

***** In BCD, a digit is usually represented by four binary (0 or 1) bits. For example the numbers 0, 1, 2, 3, 4 in the decimal system are written as 0000, 0001, 0010, 0011 and 0100. 1000 would represent eight.

****** It is a character encoding in terms of numbers based on English alphabet since the computer can only understand numbers.

(vii) ***Amplification:*** It is the process of *increasing the amplitude* (and consequently the strength) of a signal using an electronic circuit called the amplifier (reference Chapter 15). Amplification is necessary to compensate for the attenuation of the signal in communication systems. The energy needed for additional signal strength is obtained from a DC power source. Amplification is done at a place between the source and the destination wherever signal strength becomes weaker than the required strength.

(viii) ***Range:*** It is the largest distance between a source and a destination up to which the signal is received with sufficient strength.

(ix) ***Bandwidth:*** Bandwidth refers to the frequency range over which equipment operates or the portion of the spectrum occupied by the signal.

(x) ***Modulation:*** The original low frequency message/information signal cannot be transmitted to long distances because of reasons given in Section 16.7. Therefore, at the transmitter, information contained in the low frequency message signal is superimposed on a high frequency wave, which acts as a carrier of the information. This process is known as modulation. As will be explained later, there are several types of modulation, abbreviated as AM, FM and PM.

(xi) ***Demodulation:*** The process of retrieval of information from the carrier wave at the receiver is termed demodulation. This is the reverse process of modulation.

(xii) ***Repeater:*** A repeater is a combination of a receiver and a transmitter. A repeater, picks up the signal from the transmitter, amplifies and retransmits it to the receiver sometimes with a change in carrier frequency. Repeaters are used to extend the range of a communication system as shown in Fig. 16.2. A communication satellite is essentially a repeater station in space.

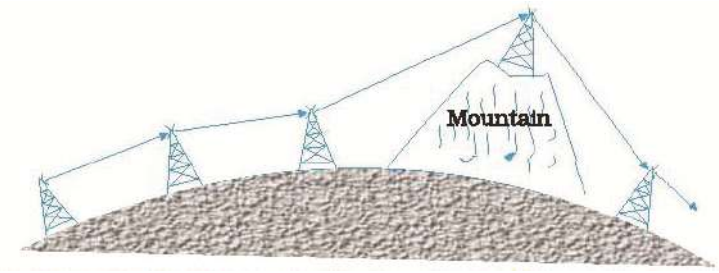


Fig. 16.2 Use of repeater station to increase the range of communication.

16.4 BANDWIDTH OF SIGNALS

In a communication system, the message signal can be voice, music, and picture or computer data. Each of these signals has different ranges of frequencies. The type of communication system needed for a given signal depends on the band of frequencies which is considered essential for the communication process.

For speech signals, frequency range 300 Hz to 3100 Hz is considered adequate. Therefore speech signal requires a bandwidth of 2800 Hz (3100 Hz – 300 Hz) for commercial telephonic communication. To transmit music, an approximate bandwidth of 20 kHz is required because of the high frequencies produced by the musical instruments. The audible range of frequencies extends from 20 Hz to 20 kHz.

Video signals for transmission of pictures require about 4.2 MHz of bandwidth. A TV signal contains both voice and picture and is usually allocated 6 MHz of bandwidth for transmission.

In the preceding paragraph, we have considered only analog signals. Digital signals are in the form of rectangular waves as shown in Fig. 16.3. One can show that this rectangular wave can be decomposed into a superposition of sinusoidal waves of frequencies ν_0, $2\nu_0$, $3\nu_0$,

$4\nu_0$ ... $n\nu_0$ where n is an integer extending to infinity and $\nu_0 = 1/T_0$. The fundamental (ν_0), fundamental (ν_0) + second harmonic ($2\nu_0$), and fundamental (ν_0) + second harmonic ($2\nu_0$) + third harmonic ($3\nu_0$), are shown in the same figure to illustrate this fact. It is clear that to reproduce the rectangular wave shape exactly we need to superimpose all the harmonics ν_0, $2\nu_0$, $3\nu_0$, $4\nu_0$..., which implies an infinite bandwidth. However, for practical purposes, the contribution from higher harmonics can be neglected, thus limiting the bandwidth. As a result, received waves are a distorted version of the transmitted one. If the bandwidth is large enough to accommodate a few harmonics, the information is not lost and the rectangular signal is more or less recovered. This is so because the higher the harmonic, less is its contribution to the wave form.

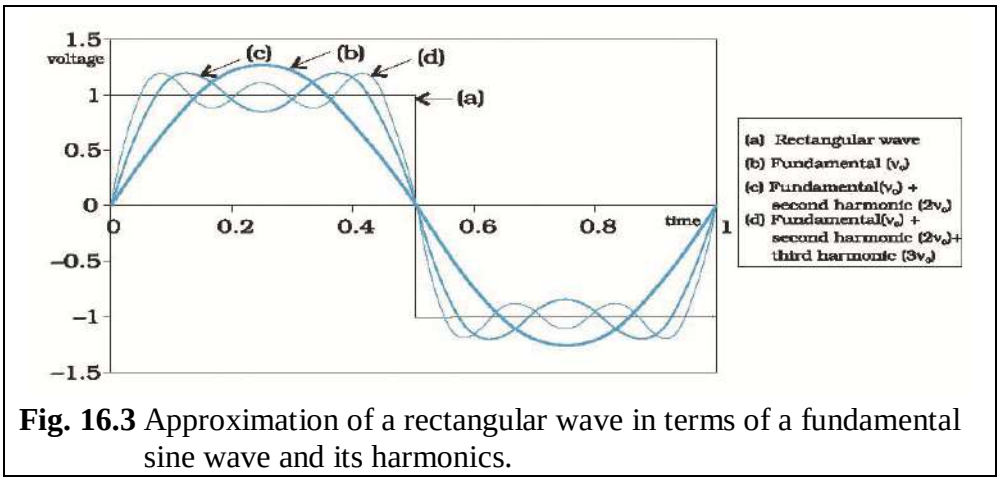


Fig. 16.3 Approximation of a rectangular wave in terms of a fundamental sine wave and its harmonics.

16.5 BANDWIDTH OF TRANSMISSION MEDIUM

Similar to message signals, different types of transmission media offer different bandwidths. The commonly used transmission media are wire, free space and fiber optic cable. Coaxial cable is a widely used wire medium, which offers a bandwidth of approximately 750 MHz. Such cables are normally operated below 18 GHz. Communication through free space using radio waves takes place over a very wide range of frequencies: from a few hundreds of kHz to a few GHz. This range of frequencies is further subdivided and allocated for various services as indicated in Table 15.2. Optical communication using fibers is performed in the frequency range of 1 THz to 1000 THz (microwaves to ultraviolet). An optical fiber can offer a transmission bandwidth in excess of 100 GHz spectrum allocations are arrived at by an international agreement. The International Telecommunication Union (ITU) administers the present system of frequency allocations.

TABLE 16.2 SOME IMPORTANT WIRELESS COMMUNICATION FREQUENCY BANDS

Service	**Frequency bands**	**Comments**
Standard AM broadcast	540-1600 kHz	
FM broadcast	88-108 MHz	
Television	54-72 MHz	VHF (very high frequencies)
	76-88 MHz	TV
	174-216 MHz	UHF (ultra high frequencies)
	420-890 MHz	TV
Cellular Mobile Radio	896-901 MHz	Mobile to base station
	840-935 MHz	Base station to mobile
Satellite Communication	5.925-6.425 GHz	Uplink
	3.7-4.2 GHz	Downlink

16.6 PROPAGATION OF ELECTROMAGNETIC WAVES

In communication using radio waves, an antenna at the transmitter radiates the Electromagnetic waves (em waves), which travel through the space and reach the receiving antenna at the other end. As the em wave travels away from the transmitter, the strength of the wave keeps on decreasing. Several factors influence the propagation of em waves and the path they follow. At this point, it is also important to understand the composition of the earth's atmosphere as it plays a vital role in the propagation of em waves. A brief discussion on some useful layers of the atmosphere is given in Table 16.3.

16.6.1 Ground wave

To radiate signals with high efficiency, the antennas should have a size comparable to the wavelength λ of the signal (at least ~ $\lambda/4$). At longer wavelengths (i.e., at lower frequencies), the antennas have large physical size and they are located on or very near to the ground. In standard AM broadcast, ground based vertical towers are generally used as transmitting antennas. For such antennas, ground has a strong influence on the propagation of the signal. The mode of propagation is called surface wave propagation and the wave glides over the surface of the earth. A wave induces current in the ground over which it passes and it is attenuated as a result of absorption of energy by the earth. The attenuation of surface waves increases very rapidly with increase in frequency. The maximum range of coverage depends on the transmitted power and frequency (less than a few MHz).

TABLE 16.3 DIFFERENT LAYERS OF ATMOSPHERE AND THEIR INTERACTION WITH THE PROPAGATING ELECTROMAGNETIC WAVES

Name of the stratum (layer)		**Approximate height over earth's surface**	**Exists during**	**Frequencies most affected**
Troposphere		10 km	Day and night	VHF (up to several GHz)
D (part of stratosphere)	PARTS OF IONOSPHERE	65-75 km	Day only	Reflects LF, absorbs MF and HF to some degree
E (part of Stratosphere)		100 km	Day only	Helps surface waves, reflects HF
F1 (Part of Mesosphere)		170-190 km	Daytime, merges with F2 at night	Partially absorbs HF waves yet allowing them to reach F_2
F2 (Thermosphere)		300 km at night, 250-400 km during daytime	Day and night	Efficiently reflects HF waves, particularly at night

16.6.2 Sky waves

In the frequency range from a few MHz up to 30 to 40 MHz, long distance communication can be achieved by ionospheric reflection of radio waves back towards the earth. This mode of propagation is called *sky wave propagation* and is used by short wave broadcast services. The ionosphere is so called because of the presence of a large number of ions or charged particles. It extends from a height of ~ 65 Km to about 400 Km above the earth's surface. Ionisation occurs due to the absorption of the ultraviolet and other high-energy radiation coming from the sun by air molecules. The ionosphere is further subdivided into several layers, the details of which are given in Table 16.3. The degree of ionisation varies with the height. The density of atmosphere decreases with height. At great heights the solar radiation is intense but there are few molecules to be ionised. Close to the earth, even though the molecular concentration is very high, the radiation intensity is low so that the

ionisation is again low. However, at some intermediate heights, there occurs a peak of ionisation density. The ionospheric layer acts as a reflector for a certain range of frequencies (3 to 30 MHz). Electromagnetic waves of frequencies higher than 30 MHz penetrate the ionosphere and escape. These phenomena are shown in the Fig. 16.4. The phenomenon of bending of em waves so that they are diverted towards the earth is similar to total internal reflection in optics.

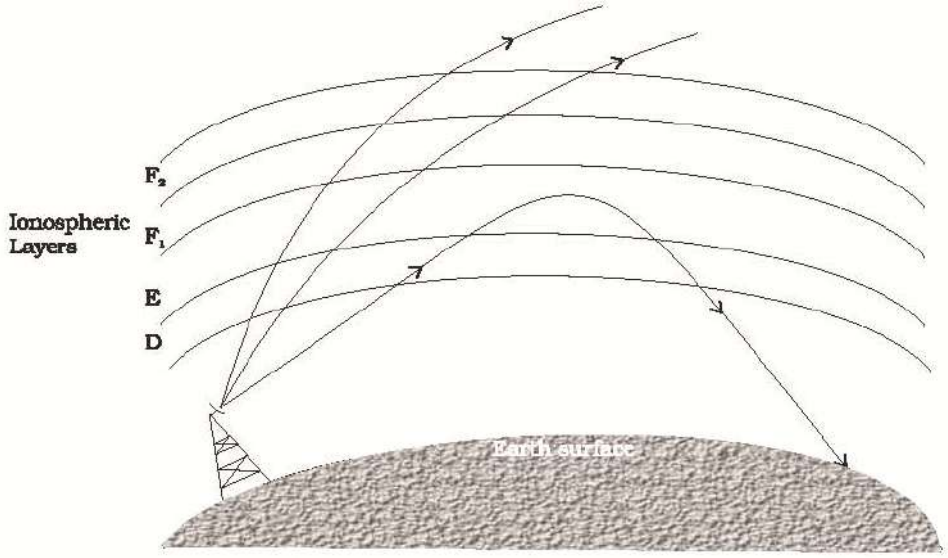


Fig.16.4 Sky wave propagation. The layer nomenclature is given in Table 16.3.

16.6.3 Space wave

Another mode of radio wave propagation is by *space waves*. A space wave travels in a straight line from transmitting antenna to the receiving antenna. Space waves are used for line-of-sight (LOS) communication as well as satellite communication. At frequencies above 40 MHz, communication is essentially limited to line-of-sight paths. At these frequencies, the antennas are relatively smaller and can be placed at heights of many wavelengths above the ground. Because of line-of-sight nature of propagation, direct waves get blocked at some point by the curvature of the earth as illustrated in Fig. 16.5. If the signal is to be received beyond the horizon then the receiving antenna must be high enough to intercept the line-of-sight waves.

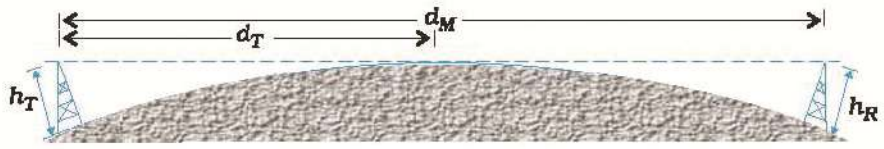


Fig. 16.5 Line of sight communication by space waves.

If the transmitting antenna is at a height h_T, then you can show that the distance to the horizon d_T is given as, where R is the radius of the earth (approximately 6400 km). d_T is also called the radio horizon of the transmitting antenna. With reference to Fig. 16.5 the maximum line-of-sight distance d_M between the two antennas having heights h_T and h_R above the earth is given by

$$d_M = \sqrt{2Rh_T} + \sqrt{2Rh_R} \qquad (16.1)$$

where h_R is the height of receiving antenna.

Television broadcast, microwave links and satellite communication are some examples of communication systems that use space wave mode of propagation. Figure 16.6 summarises the various modes of wave propagation discussed so far.

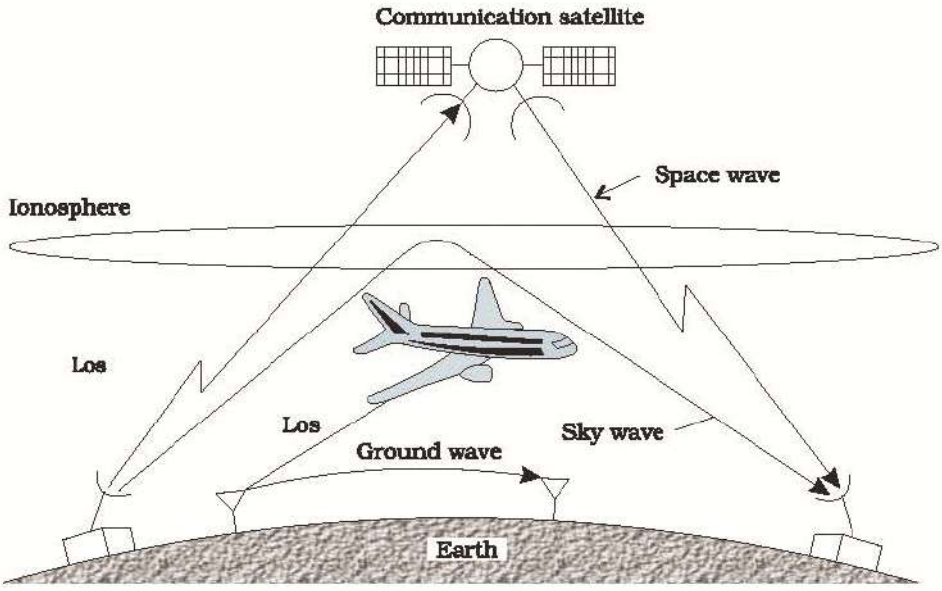


Fig. 16.6 Various propagation modes for em waves.

16.7 MODULATION AND ITS NECESSITY

As already mentioned, the purpose of a communication system is to transmit information or message signals. Message signals are also called *baseband signals*, which essentially designate the band of frequencies representing the original signal, as delivered by the source of information. No signal, in general, is a single frequency sinusoid, but it spreads over a range of frequencies called the signal *bandwidth*. Suppose we wish to transmit an electronic signal in the audio frequency (AF) range (baseband signal frequency less than 20 kHz) over a long distance directly. Let us find what factors prevent us from doing so and how we overcome these factors,

16.7.1 Size of the antenna or aerial

For transmitting a signal, we need an antenna or an aerial. This antenna should have a size comparable to the wavelength of the signal (at least $\lambda/4$ in dimension) so that the antenna properly senses the time variation of the signal. For an electromagnetic wave of frequency 20 kHz, the wavelength λ is 15 km. Obviously, such a long antenna is not possible to construct and operate. Hence direct transmission of such baseband signals is not practical. We can obtain transmission with reasonable antenna lengths if transmission frequency is high (for example, if v is 1 MHz, then λ is 300 m). Therefore, there is a need of *translating the information contained in our original low frequency baseband signal into high or radio frequencies before transmission.*

16.7.2 Effective power radiated by an antenna

A theoretical study of radiation from a linear antenna (length l) shows that the power radiated is proportional to $(l/\lambda)^2$. This implies that for the same antenna length, the power radiated increases with decreasing λ, i.e., increasing frequency. Hence, the effective power radiated by a long wavelength baseband signal would be small. For a good transmission, we need high powers and hence this also points out to *the need* of using high frequency transmission.

16.7.3 Mixing up of signals from different transmitters

Another important argument against transmitting baseband signals directly is more *practical* in nature. Suppose many people are talking at the same time or many transmitters are transmitting baseband information signals simultaneously. All these signals will get mixed up and there is no simple way to distinguish between them. This points out towards a possible solution by using communication at high frequencies and allotting a *band* of frequencies to each message signal for its transmission.

The above arguments suggest that there is a *need for translating the original low frequency baseband message or information signal into high frequency wave before transmission such that the translated signal continues to possess the information contained in the original signal*. In doing so, we take the help of a high frequency signal, known as the *carrier wave*, and a process known as modulation which attaches information to it. The carrier wave may be continuous (sinusoidal) or in the form of pulses as shown in Fig. 16.7.

A sinusoidal carrier wave can be represented as

$$c(t) = A_c \sin(\omega_c t + \phi) \qquad (16.2)$$

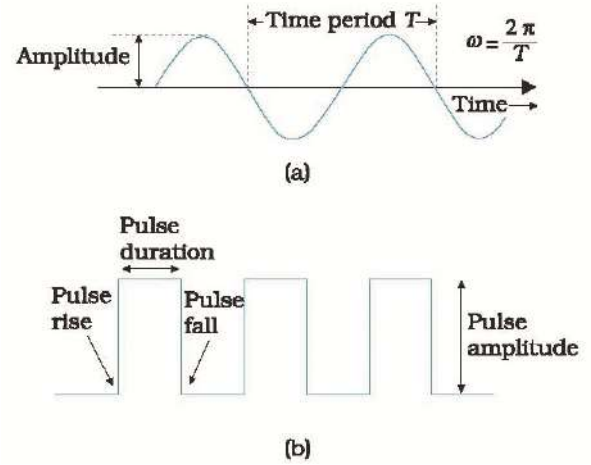


Fig. 16.7 (a) Sinusoidal, and (b) pulse shaped signals.

where $c(t)$ is the signal strength (voltage or current), A_c is the amplitude, ω_c (= $2\pi \nu_c$) is the angular frequency and ϕ is the initial phase of the carrier wave. During the process of modulation, any of the three parameters, *viz* A_c, ω_c and ϕ, of the carrier wave can be controlled by the message or information signal. This results in three types of modulation: (i) Amplitude modulation (AM), (ii) Frequency modulation (FM) and (iii) Phase modulation (PM), as shown in Fig. 16.8.

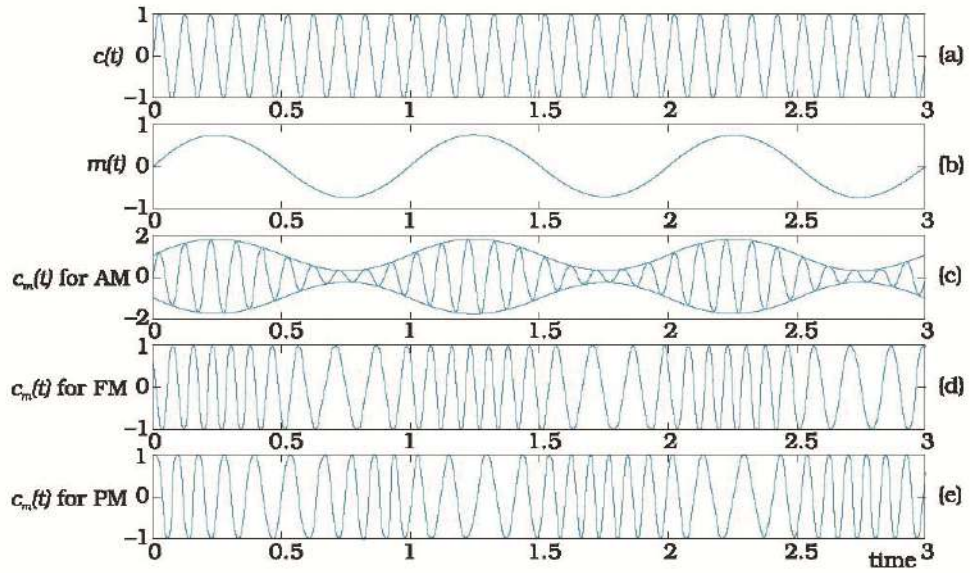


Fig. 16.8 Modulation of a carrier wave: (a) a sinusoidal carrier wave; (b) a modulating signal; (c) amplitude modulation; (d) frequency modulation; and (e) phase modulation.

Similarly, the significant characteristics of a pulse are: pulse amplitude, pulse duration or pulse Width, and pulse position (denoting the time of *rise* or *fall* of the pulse amplitude) as shown in Fig. 16.7(b). Hence, different types of pulse modulation are: (a) pulse amplitude modulation (PAM), (b) pulse duration modulation (PDM) or pulse width modulation (PWM), and (c) pulse position modulation (PPM).

ADDITIONAL INFORMATION

The Internet

It is a system with billions of users worldwide. It permits communication and sharing of all types of information between any two or more computers connected through a large and complex network. It was started in 1960's and opened for public use in 1990's. With the passage of time it has witnessed tremendous growth and it is still expanding its reach. Its applications include

(i) *E-mail* – It permits exchange of text/graphic material using email software. We can write a letter and send it to the recipient through ISP's (Internet Service Providers) who work like the dispatching and receiving post offices.

(ii) *File transfer* – A FTP (File Transfer Programmes) allows transfer of files/software from one computer to another connected to the Internet.

(iii) *World Wide Web (WWW)* – Computers that store specific information for sharing with others provides *websites* either directly or through web service providers. Government departments, companies, NGO's (Non-Government Organisations) and individuals can post information about their activities for restricted or free use on their websites. This information becomes accessible to the users. Several search engines like Google, Yahoo! etc. help us in finding information by listing the related websites. *Hypertext* is a powerful feature of the web that automatically links relevant information from one page on the web to another using *HTML (hypertext markup language)*.

(iv) *E-commerce* – Use of the Internet to promote business using electronic means such as using credit cards is called E-commerce. Customers view images and receive all the information about various products or services of companies through their websites. They can do *on-line shopping* from home/office. Goods are dispatched or services are provided by the company through mail/courier.

(v) *Chat* – Real time conversation among people with common interests through typed messages is called chat. Everyone belonging to the *chat group* gets the message instantaneously and can respond rapidly.

Facsimile (FAX)

It scans the contents of a document (as an image, not text) to create electronic signals. These signals are then sent to the destination (another FAX machine) in an orderly manner using telephone lines. At the destination, the signals are reconverted into a replica of the original document. Note that FAX provides image of a static document unlike the image provided by television of objects that might be dynamic.

Mobile telephony

The concept of mobile telephony was developed first in 1970's and it was fully implemented in the following decade. The central concept of this system is to divide the service area into a suitable number of *cells* centred on an office called *MTSO (Mobile Telephone Switching Office)*. Each cell contains a low-power transmitter called a *base station* and caters to a large number of mobile receivers (popularly called cell phones). Each cell could have a service area of a few square kilometers or even less depending upon the number of customers. When a mobile receiver crosses the coverage area of one base station, it is necessary for the mobile user to be transferred to another base station. This procedure is called *handover* or *handoff*. This process is carried out very rapidly, to the extent that the consumer does not even notice it. Mobile telephones operate typically in the UHF range of frequencies (about 800-950 MHz).

SUMMARY

1. Electronic communication refers to the faithful transfer of information or message (available in the form of electrical voltage and current) from one point to another point.
2. Transmitter, transmission channel and receiver are three basic units of a communication system.
3. Two important forms of communication system are: Analog and Digital. The information to be transmitted is generally in continuous waveform for the former while for the latter it has only discrete or quantised levels.
4. Every message signal occupies a range of frequencies. The bandwidth of a message signal refers to the band of frequencies, which are necessary for satisfactory transmission of the information contained in the signal. Similarly, any practical communication system permits transmission of a range of frequencies only, which is referred to as the bandwidth of the system.
5. Low frequencies cannot be transmitted to long distances. Therefore, they are superimposed on a high frequency carrier signal by a process known as modulation.
6. In modulation, some characteristic of the carrier signal like amplitude, frequency or phase varies in accordance with the modulating or message signal. Correspondingly, they are called Amplitude Modulated (AM), Frequency Modulated (FM) or Phase Modulated (PM) waves.
7. Pulse modulation could be classified as: Pulse Amplitude Modulation (PAM), Pulse Duration Modulation (PDM) or Pulse Width Modulation (PWM) and Pulse Position Modulation (PPM).
8. For transmission over long distances, signals are radiated into space using devices called antennas. The radiated signals propagate as electromagnetic waves and the mode of propagation is influenced by the presence of the earth and its atmosphere. Near the

surface of the earth, electromagnetic waves propagate as surface waves. Surface wave propagation is useful up to a few MHz frequencies.

9. Long distance communication between two points on the earth is achieved through reflection of electromagnetic waves by ionosphere. Such waves are called sky waves. Sky wave propagation takes place up to frequency of about 30 MHz. Above this frequency, electromagnetic waves essentially propagate as space waves. Space waves are used for line-of-sight communication and satellite communication.
10. If an antenna radiates electromagnetic waves from a height h_T, then the range d_T is given by $\sqrt{2Rh_T}$ where R is the radius of the earth.
11. Amplitude modulated signal contains frequencies $(\omega_c - \omega_m)$, ω_c and $(\omega_c + \omega_m)$.

VERY SHORT ANSWER QUESTIONS

1. What are the basic blocks of a communication system?
2. What is 'World Wide Web' (www)?
3. Mention the frequency range of speech signals.
4. What is sky wave propagation?
5. Mention the various parts of the ionosphere.
6. Define modulation. Why is it necessary?
7. Mention the basic methods of modulation.
8. Which type of communication is employed in Mobile Phones?

SHORT ANSWER QUESTIONS

1. Draw the block diagram of a generalised communication system and explain it briefly.
2. What is ground wave? When is it used for communication?
3. What are sky waves? Explain Sky wave propagation. Briefly.
4. What is Space Wave Communication? Explain.

Chemistry

Chemistry-II

CONTENTS

CHAPTER 1

THE SOLID STATE

The vast majority of solid substances like high temperature superconductors, biocompatible plastics, silicon chips, etc. are destined to play an ever expanding role in future development of science.

We are mostly surrounded by solids; we use them more often than liquids and gases. For different applications we need solids with widely different properties. These properties depend upon the nature of constituent particles and the binding forces operating between them. Therefore, study of the structure of solids is important. The correlation between structure and properties helps in discovering new solid materials with desired properties like high temperature superconductors, magnetic materials, and biodegradable polymers for packaging, bio-compliant solids for surgical implants, etc.

From our earlier studies, we know that liquids and gases are called *fluids* because of their ability to flow. The fluidity in both of these states is due to the fact that the molecules are free to move about. On the contrary, the constituent particles in solids have fixed positions and can only oscillate about their mean positions. This explains the rigidity in solids. In crystalline solids, the constituent particles are arranged in regular patterns.

In this Unit, we shall discuss different possible arrangements of particles resulting in several types of structures. The correlation between the nature of interactions within the constituent particles and several properties of solids will also be explored. How these properties get modified due to the structural imperfections or by the presence of impurities in minute amounts would also be discussed.

1.1 General Characteristics of Solid State

In Class XI you have learnt that matter can exist in three states namely, solid, liquid and gas. Under a given set of conditions of temperature and pressure, which of these would be the most stable state of a given substance depends upon the net effect of two opposing factors. Intermolecular forces tend to keep the molecules (or atoms or ions) closer, whereas thermal energy tends to keep them apart by making them move faster. At sufficiently low temperature, the thermal energy is low and intermolecular forces bring them so close that they cling to one another and occupy fixed positions. These can still oscillate about their mean positions and the substance exists in solid state. The following are the characteristic properties of the solid state:

(i) They have definite mass, volume and shape.

(ii) Intermolecular distances are short.

(iii) Intermolecular forces are strong.

(iv) Their constituent particles (atoms, molecules or ions) have fixed positions and can only oscillate about their mean positions.

(v) They are incompressible and rigid.

1.2 Amorphous and Crystalline Solids

Solids can be classified as *crystalline* or *amorphous* on the basis of the nature of order present in the arrangement of their constituent particles. A crystalline solid usually consists of a large number of small crystals, each of them having a definite characteristic geometrical shape. In a crystal, the arrangement of constituent particles (atoms, molecules or ions) is ordered. It has long range order which means that there is a regular pattern of arrangement of particles which repeats itself periodically over the entire crystal. Sodium chloride and quartz

are typical examples of crystalline solids. An amorphous solid (Greek *amorphos* = no form) consists of particles of irregular shape. The arrangement of constituent particles (atoms, molecules or ions) in such a solid has only *short range order*. In such an arrangement, a regular and periodically repeating pattern is observed over short distances only. Such portions are scattered and in between the arrangement is disordered. The structures of quartz (crystalline) and quartz glass (amorphous) are shown in the following figures 1.1 (a) and (b) respectively. While the two structures are almost identical, yet in the case of amorphous quartz glass there is no *long range order*. The structure of amorphous solids is similar to that of liquids. Glass, rubber and plastics are typical examples of amorphous solids. Due to the differences in the arrangement of the constituent particles, the two types of solids differ in their properties.

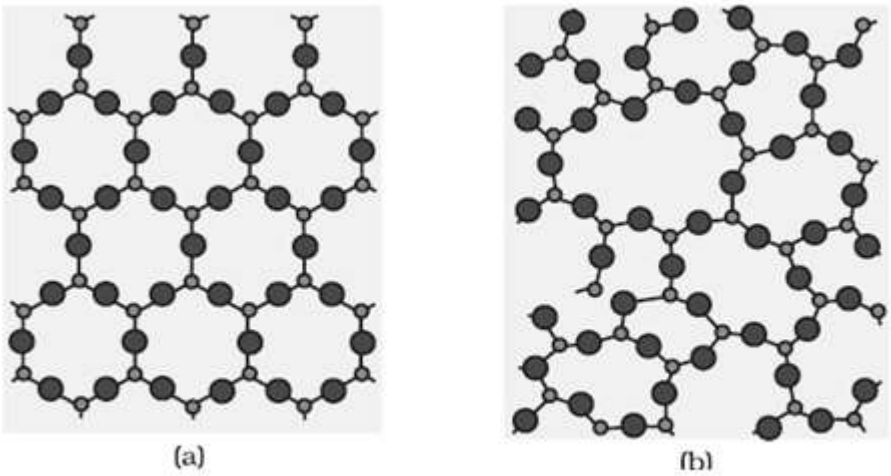


Fig. 1.1 Two dimensional structure of (a) quartz and (b) quartz glass

Crystalline solids have a sharp melting point. On the other hand, amorphous solids soften over a range of temperature and can be moulded and blown into various shapes. On heating they become crystalline at some temperature. Some glass objects from ancient civilisations are found to become milky in appearance because of some crystallisation. Like liquids, amorphous solids have a tendency to flow, though very slowly. Therefore, sometimes these are called *pseudo solids* or *super cooled liquids*. Glass panes fixed to windows or doors of old buildings are invariably found to be slightly thicker at the bottom than at the top. This is because the glass flows down very slowly and makes the bottom portion slightly thicker.

Crystalline solids are *anisotropic* in nature, that is, some of their physical properties like electrical resistance or refractive index show different values when measured along different directions in the same crystals. This arises from different arrangement of particles in different directions. This is illustrated in figure 1.2. Since the arrangement of particles is different along different directions, the value of same physical property is found to be different along each direction.

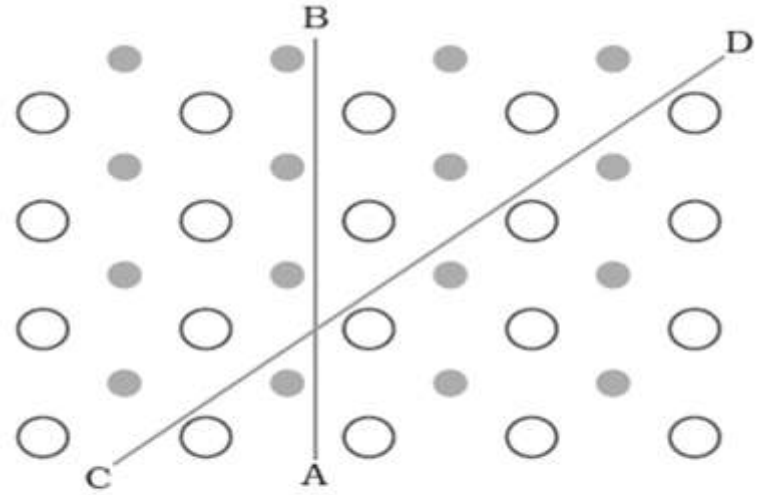


Fig. 1.2 Anisotropy in crystals is due to different arrangement of particles along different directions

Amorphous solids on the other hand are *isotropic* in nature. It is because there is no *long range* order in them and arrangement is irregular along all the directions. Therefore, value of any physical property would be same along any direction. These differences are summarized in the following Table1.1.

Table 1.1. Distinction between Crystalline and Amorphous Solids

Property	Crystalline solids	Amorphous solids
Shape	Definite characteristic geometrical shape	Irregular shape
Melting point	Melt at a sharp and characteristic temperature	Gradually soften over a range of temperature
Cleavage property	When cut with a sharp edged tool, they split into two pieces and the newly generated surfaces are plain and smooth	When cut with a sharp edged tool, they cut into two pieces with irregular surfaces
Heat of fusion	They have a definite and characteristic heat of fusion	They do not have definite heat of fusion
Anisotropy	Anisotropic in nature	Isotropic in nature
Nature arrangement of constituent particles	True solids Long range order	Pseudo solids or super cooled liquids Only short range order.

Amorphous solids are useful materials. Glass, rubber and plastics find many applications in our daily lives. Amorphous silicon is one of the best photovoltaic materials available for conversion of sunlight into electricity.

1.3 Classification of Crystalline Solids

In Section 1.2, we have learnt about amorphous substances and that they have only short range order. However, most of the solid substances are crystalline in nature. For

example, all the metallic elements like iron, copper and silver; non – metallic elements like sulphur, phosphorus and iodine and compounds like sodium chloride, zinc sulphide and naphthalene form crystalline solids.

Crystalline solids can be classified on the basis of nature of intermolecular forces operating in them into four categories viz., molecular, ionic, metallic and covalent solids. Let us now learn about these categories.

1.3.1Molecular Solids

Molecules are the constituent particles of molecular solids. These are further sub divided into the following categories:

(i) ***Non polar Molecular Solids:*** They comprise of either atoms, for example, argon and helium or the molecules formed by non polar covalent bonds for example H_2, Cl_2 and I_2. In these solids, the atoms or molecules are held by weak dispersion forces or London forces about which you have learnt in Class XI. These solids are soft and non-conductors of electricity. They have low melting points and are usually in liquid or gaseous state at room temperature and pressure.

(ii) ***Polar Molecular Solids*:** The molecules of substances like HCl, SO_2, *etc.* are formed by polar covalent bonds. The molecules in such solids are held together by relatively stronger dipole-dipole interactions. These solids are soft and non-conductors of electricity. Their melting points are higher than those of non polar molecular solids yet most of these are gases or liquids under room temperature and pressure. Solid SO_2 and solid NH_3 are some examples of such solids.

(iii) ***Hydrogen Bonded Molecular Solids*:** The molecules of such solids contain polar covalent bonds between H and F, O or N atoms. Strong hydrogen bonding binds molecules of such solids like H_2O (ice). They are non-conductors of electricity. Generally they are volatile liquids or soft solids under room temperature and pressure.

1.3.2 Ionic Solids

Ions are the constituent particles of ionic solids. Such solids are formed by the three dimensional arrangements of cations and anions bound by strong coulombic (electrostatic) forces. These solids are hard and brittle in nature. They have high melting and boiling points. Since the ions are not free to move about, they are electrical insulators in the solid state. However, in the molten state or when dissolved in water, the ions become free to move about and they conduct electricity.

1.3.3 Metallic Solids

Metals are orderly collection of positive ions surrounded by and held together by a sea of free electrons. These electrons are mobile and are evenly spread out throughout the crystal. Each metal atom contributes one or more electrons towards this sea of mobile electrons. These free and mobile electrons are responsible for high electrical and thermal conductivity of metals. When an electric field is applied, these electrons flow through the network of positive ions. Similarly, when heat is supplied to one portion of a metal, the thermal energy is uniformly spread throughout by free electrons. Another important characteristic of metals is their lustre and colour in certain cases. This is also due to the presence of free electrons in them. Metals are highly malleable and ductile.

1.3.4 Covalent or Network Solids

A wide variety of crystalline solids of non-metals result from the formation of covalent bonds between adjacent atoms throughout the crystal. They are also called **giant molecules.** Covalent bonds are strong and directional in nature, therefore atoms are held very strongly at their positions. Such solids are very hard and brittle. They have extremely high

melting points and may even decompose before melting. They are insulators and do not conduct electricity. Diamond (figure 1.3) and silicon carbide are typical examples of such solids. Graphite is soft and a conductor of electricity. Its exceptional properties are due to its typical structure (figure 1.4). Carbon atoms are arranged in different layers and each atom is covalently bonded to three of its neighbouring atoms in the same layer. The fourth valence electron of each atom is present between different layers and is free to move about. These free electrons make graphite a good conductor of electricity. Different layers can slide one over the other. This makes graphite a soft solid and a good solid lubricant.

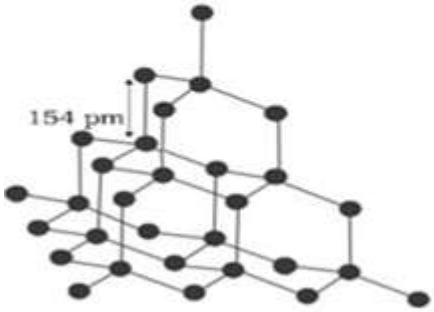


Fig. 1.3 Network structure of diamond

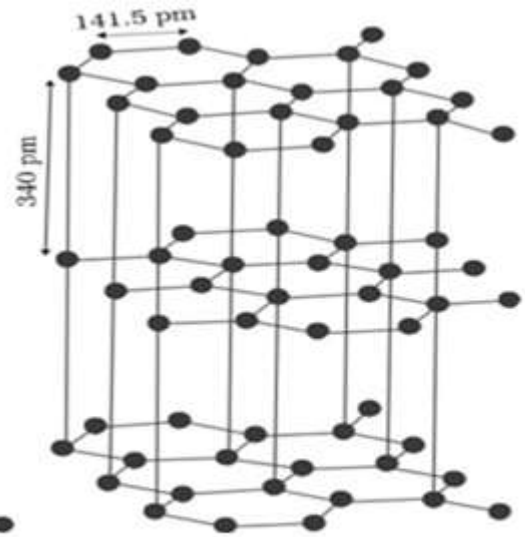


Fig. 1.4 Structure of graphine

1.4 Crystal Lattices and Unit Cells

The main characteristic of crystalline solids is a regular and repeating pattern of constituent particles. If the three dimensional arrangement of constituent particles in a crystal is represented diagrammatically, in which each particle is depicted as a point, the arrangement is called *crystal lattice*. Thus, a regular three dimensional arrangement of points in space is called a **crystal lattice**. A portion of a crystal lattice is shown in the following figure 1.5. The different properties of the four types of solids are listed in Table 1.2.

Table 1.2. Different Types of Solids

Type of Solid	Constituent Particles	Bonding/ Attractive Forces	Examples	Physical Nature	Electrical Conductivity	Melting Point
(1) Molecular solids						
(i) Non polar	Molecules	Dispersion or London forces	Ar, CCl_4, H_2, I_2, CO_2	Soft	Insulator	Very low
(ii) Polar		Dipole-dipole interactions	HCl, SO_2	Soft	Insulator	Low
(iii) Hydrogen bonded		Hydrogen bonding	H_2O (ice)	Hard	Insulator	Low
(2) Ionic solids	Ions	Coulombic or electrostatic	NaCl, MgO, ZnS, CaF_2	Hard but brittle	Insulators in solid state but conductors in molten state and in aqueous solutions	High
(3) Metallic solids	Positive ions in a sea of delocalised electrons	Metallic bonding	Fe, Cu, Ag,Mg	Hard but malleable and ductile	Conductors in solid state as well as in molten state	Fairly high
(4) Covalent or network solids	Atoms	Covalent bonding.	SiO_2(quartz), SiC, C (diamond), AlN	Hard	Insulators	Very high
			$C_{(graphite)}$	Soft	Conductor (exception)	

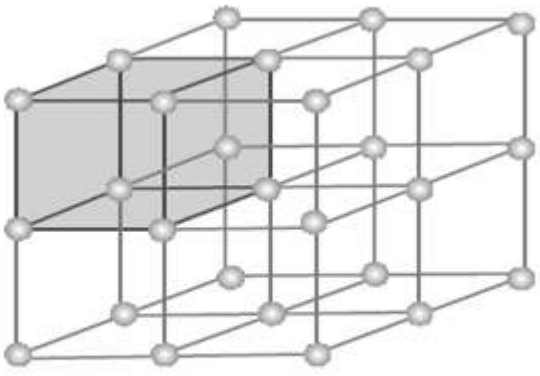


Fig. 1.5 A portion of a three dimensional cubic latice and its unit cell

There are only 14 possible three dimensional lattices. These are called **Bravais Lattices** (after the French mathematician who first described them). The following are the characteristics of a crystal lattice:

(a) Each point in a lattice is called lattice point or lattice site.
(b) Each point in a crystal lattice represents one constituent particle which may be an atom, a molecule (group of atoms) or an ion.
(c) Lattice points are joined by straight lines to bring out the geometry of the lattice.

Unit cell is the smallest portion of a crystal lattice which, when repeated in different directions, generates the entire lattice. A unit cell is characterised by:

(i) its dimensions along the three edges, *a*, *b* and *c*. These edges may or may not be mutually perpendicular.
(ii) angles between the edges, α (between *b* and *c*) β (between *a* and *c*) and γ (between *a* and *b*).

Thus, a unit cell is characterised by six parameters, *a, b, c*, α, β and γ. These parameters of a typical unit cell .

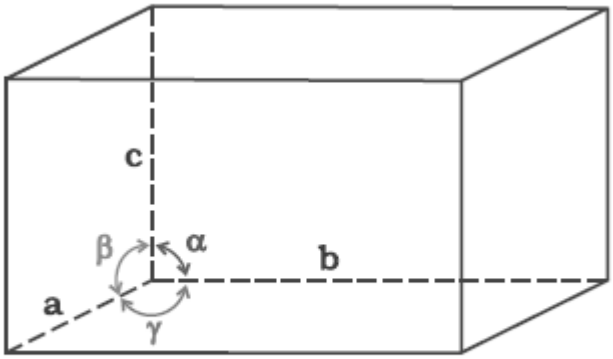


Fig.1.6. Illustration of parameters of a unit cell

1.4.1 Primitive and Centred Unit Cells

Unit cells can be broadly divided into two categories, primitive and centred unit cells.

(a) Primitive Unit Cells: When constituent particles are present only on the corner positions of a unit cell, it is called as **primitive unit cell**.

(b) Centred Unit Cells: When a unit cell contains one or more constituent particles present at positions other than corners in addition to those at corners, it is called a **centred unit cell.** Centred unit cells are of three types:

(i) ***Body-Centred Unit Cells:*** Such a unit cell contains one constituent particle (atom, molecule or ion) at its body-centre besides the ones that are at its corners.
(ii) ***Face-Centred Unit Cells:*** Such a unit cell contains one constituent particle present at the centre of each face, besides the ones that are at its corners.
(iii) ***End-Centred Unit Cells:*** In such a unit cell, one constituent particle is present at the centre of any two opposite faces besides the ones present at its corners. In all, there are seven types of primitive unit cells (figure 1.7).

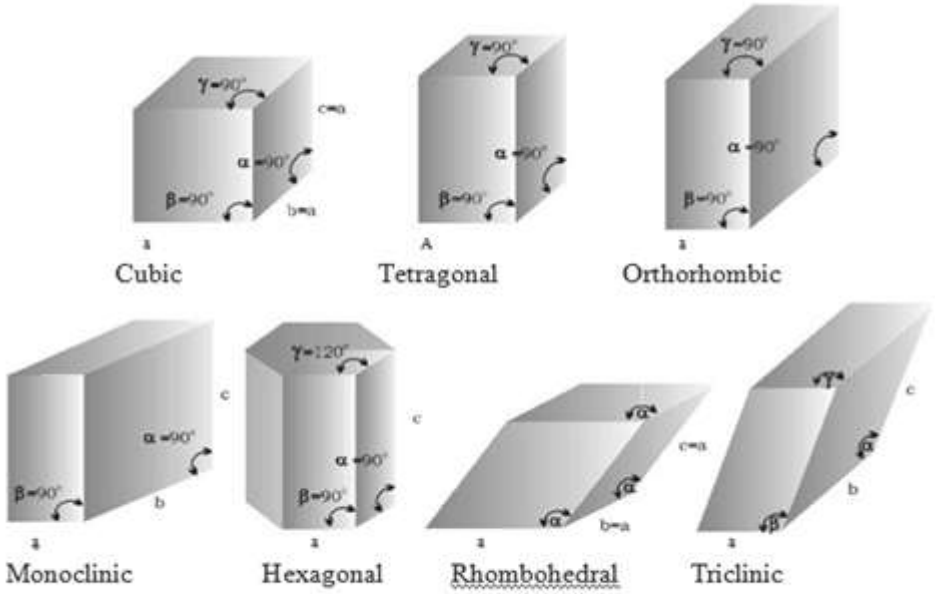


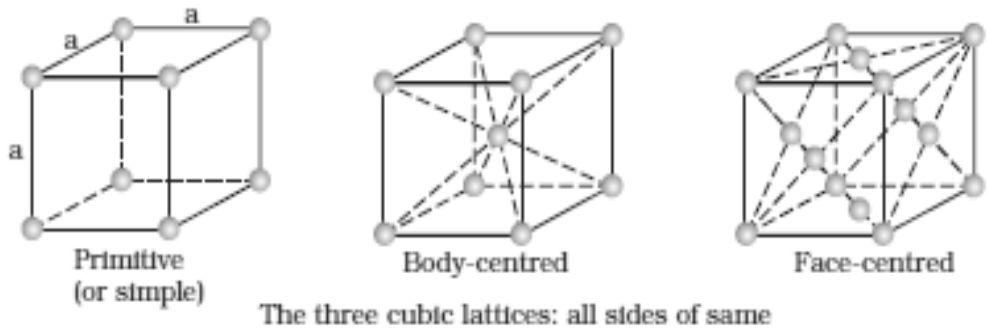
Fig. 1.7 Seven primitive unit cells in crystals

Their characteristics along with the centred unit cells they can form have been listed in Table 1.3.

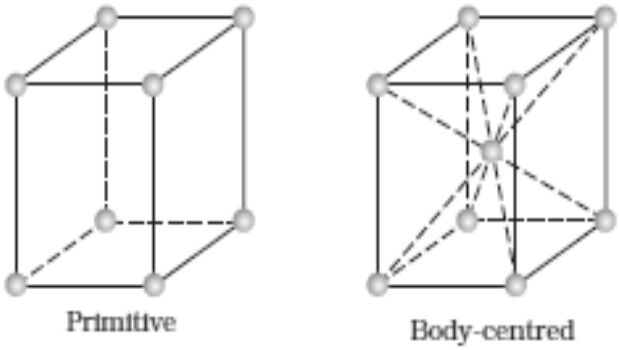
Table 1.3. Seven Primitive Unit Cells and their Possible Variations as Centred Unit Cells

Crystal system	Possible Variations	Axial distances or edge lengths	Axial angles	Examples
Cubic	Primitive,	$a = b = c$	$\alpha = \beta = \gamma = 90°$	NaCl, Zinc blende,
	Body-centred, Face-centred			Cu
Tetragonal	Primitive, Body-centred	$a = b \neq c$	$\alpha = \beta = \gamma = 90°$	White tin, SnO_2, TiO_2, $CaSO_4$
Orthorhombic	Primitive,	$a \neq b \neq c$	$\alpha = \beta = \gamma = 90°$	Rhombic sulphur,
Hexagonal	Body-centred, Face-centred, End-centred Primitive	$a = b \neq c$	$\alpha = \beta = 90°$	KNO_3, $BaSO_4$ Graphite, ZnO,CdS,
Rhombohedral or Trigonal	Primitive	$a = b = c$	$\gamma = 120°$ $\alpha = \beta = \gamma \neq 90°$	Calcite ($CaCO_3$), HgS (cinnabar)

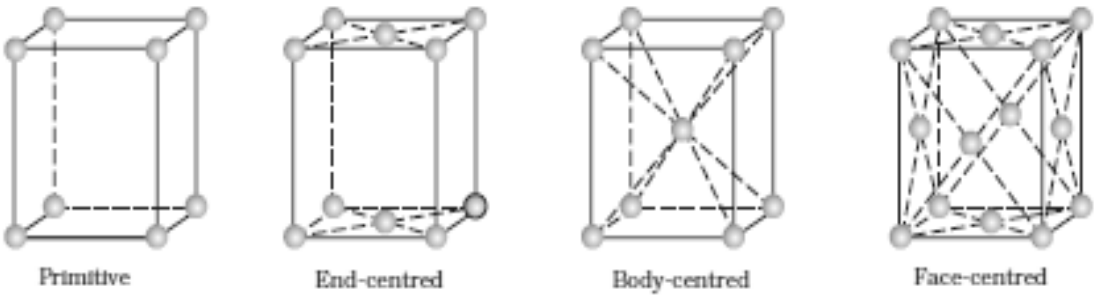
Unit Cells of 14 Types of Bravais Lattices



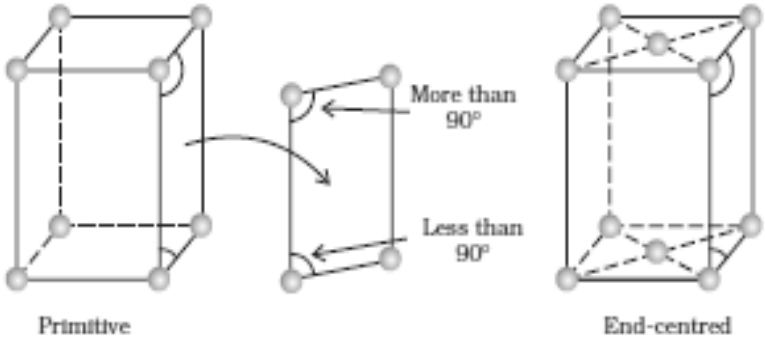
The three cubic lattices: all sides of same length, angles between faces all 90°



The two tetragonal: one side different in length to the other, two angles between faces all 90°



The four orthorhombic lattices: unequal sides, angles between faces all 90°



The two monoclinic lattices: unequal sides, two faces have angles different to 90°

1.5 Number of Atoms in a Unit Cell

1.5.1 Primitive Cubic Unit

Cell Primitive cubic unit cell has atoms only at its corner. Each atom at a corner is shared between eight adjacent unit cells as shown in Fig. 1.8, four unit cells in the same layer and four unit cells of the upper (or lower) layer. Therefore, only $1/8^{th}$ of an atom (or molecule or ion) actually belongs to a particular unit cell.

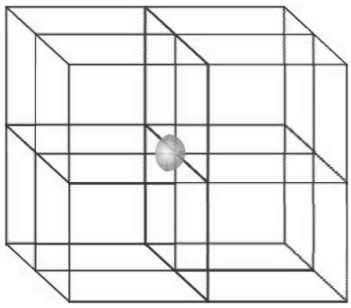


Fig.1.8 In a simple cubic unit cell, each corner atom is shared between 8 unit cells.

In a primitive cubic unit cell has been depicted in three different ways. Each small sphere in figure 1.9 (a) represents only the centre of the particle occupying that position and not its actual size. Such structures are called *open structures*. The arrangement of particles is easier to follow in open structures. Figure 1.9 (b) depicts space-filling representation of the unit cell with actual particle size and figure 1.9 (c) shows the actual portions of different atoms present in a cubic unit cell In all, since each cubic unit cell has 8 atoms on its corners, the total number of atoms in one unit cell is $8 \times \frac{1}{8} = 1$ atom.

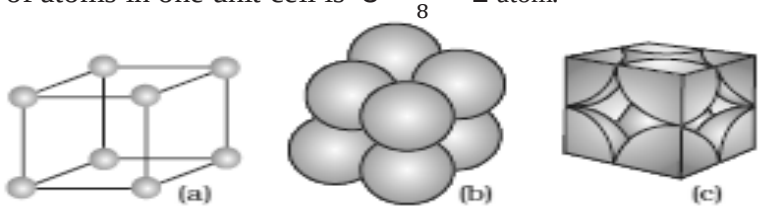


Fig. 1.9 A primitive cubic unit cell (a) open structure (b) space-filling structure (c) actual portions of atoms belonging to one unit cell.

1.5.2 Body-Centred Cubic Unit Cell

A body-centred cubic (*bcc*) unit cell has an atom at each of its corners and also one atom at its body centre. Figure 1.10 depicts (a) open structure (b) space filling model and (c) the unit cell with portions of atoms actually belonging to it. It can be seen that the atom at the

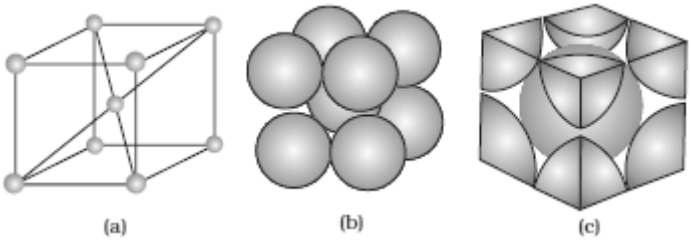


Fig.1.10 A body-centred cubic unit cell (a) open structure (b) space filling structure (c) actual portions of atoms belonging to one unit cell body centre wholly belongs to the unit cell in which it is present. Thus in a body-centered cubic (bcc) unit cell:

(i) 8 corners × $\frac{1}{8}$ per corner atom $= 8 \times \frac{1}{8}$ = 1 atom

(ii) 1 body centre atom = 1 × 1 = 1 atom

∴ Total number of atoms per unit cell = 2 atoms

1.5.3 Face – Centred Cubic Unit Cell

A face-centred cubic (*fcc*) unit cell contains atoms at all the corners and at the centre of all the faces of the cube. It can be seen in figure. 1.11 that each atom located at the face-centre is shared between two adjacent unit cells and only 1/2 of each atom belongs to a unit cell. The following figure 1.12 depicts (a) open structure (b) space-filling model and (c) the unit cell with portions of atoms actually belonging to it. Thus, in a face-centred cubic (*fcc*) unit cell:

(i) 8 corners atoms × $\frac{1}{8}$ atom per unit cell $= 8 \times \frac{1}{8}$ = 1 atom

(ii) 6 face-centred atoms × $\frac{1}{2}$ atom per unit cell = 6 × $\frac{1}{2}$ = 3 atoms

∴ Total number of atoms per unit cell = 4 atoms

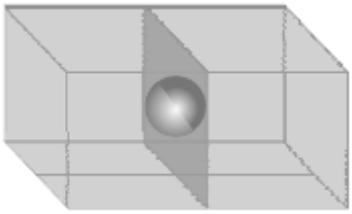


Fig.1.11 An atom at face centre of unit cell is shared between 2 unit cells

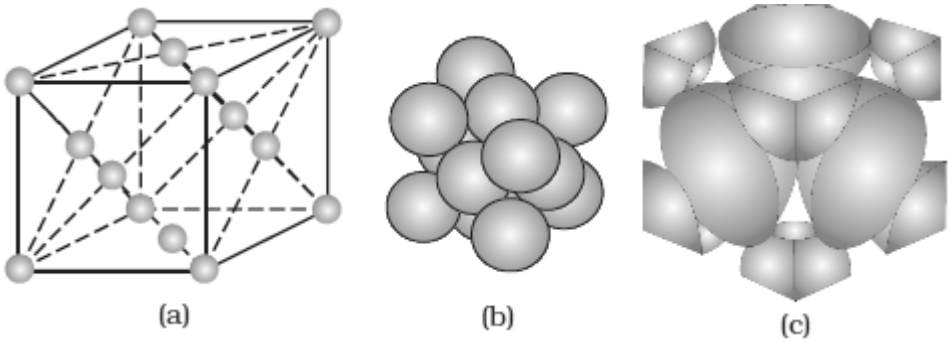


Fig. 1.12 A face-centred cubic unit cell (a) open structure (b) space filling structure (c) actual portions of atoms belonging to one unit cell.

1.6 Close Packed Structures

In solids, the constituent particles are close-packed, leaving the minimum vacant space. Let us consider the constituent particles as identical hard spheres and build up the three dimensional structure in three steps.

(a) Close Packing in One Dimension

There is only one way of arranging spheres in a one dimensional close packed structure, that is to arrange them in a row and touching each other.

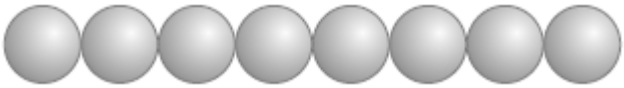


Fig. 1.13 Close packing of spheres in one dimension.

In this arrangement, each sphere is in contact with two of its neighbours. The number of nearest neighbours of a particle is called its **coordination number**. Thus, in one dimensional close packed arrangement, the coordination number is 2.

(b) Close Packing in Two Dimensions

Two dimensional close packed structure can be generated by stacking (placing) the rows of close packed spheres. This can be done in two different ways.

(i) The second row may be placed in contact with the first one such that the spheres of the second row are exactly above those of the first row. The spheres of the two rows are aligned horizontally as well as vertically. If we call the first row as 'A' type row, the second row being exactly the same as the first one, is also of 'A' type. Similarly, we may place more rows to obtain AAA type of arrangement as shown in figure 1.14 as follows.

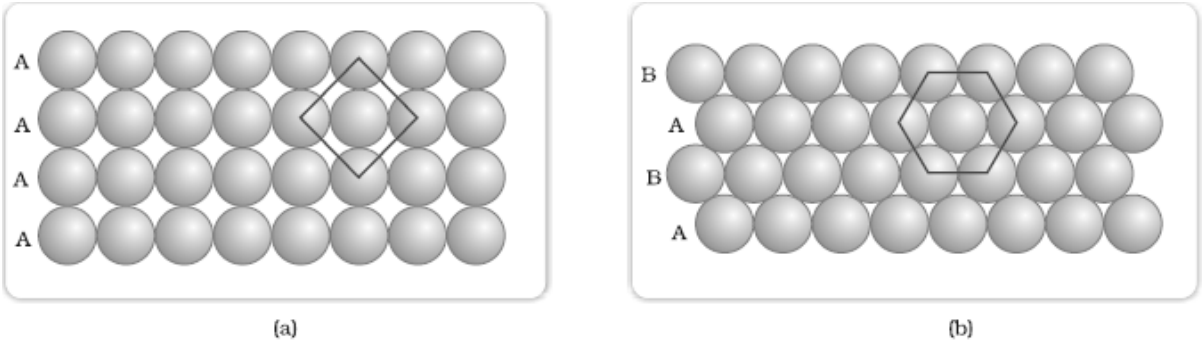


Fig.1.14 (a) Square close packing (b) hexagonal close packing of spheres in two dimensions

In this arrangement, each sphere is in contact with four of its neighbours. Thus, the two dimensional coordination number is 4. Also, if the centres of these 4 immediate neighbouring spheres are joined, a square is formed. Hence this packing is called **square close packing in two dimensions**.

(ii) The second row may be placed above the first one in a staggered manner such that its spheres fit in the depressions of the first row. If the arrangement of spheres in the first row is called 'A' type, the one in the second row is different and may be called 'B' type. When the third row is placed adjacent to the second in staggered manner, its spheres are aligned with those of the first layer. Hence this layer is also of 'A' type. The spheres of similarly placed fourth row will be aligned with those of the second row ('B' type). Hence this arrangement is of ABAB type. In this arrangement there is less free space and this packing is more efficient than the square close packing. Each sphere is in contact with six of its neighbours and the two dimensional coordination number is 6. The centres of these six spheres are at the corners of a regular hexagon hence this packing is called **two dimensional hexagonal close packing.** (b) that in this layer there are some voids (empty spaces). These are triangular in shape. The triangular voids are of two different types. In one row, the apex of the triangles is pointing upwards and in the next layer downwards.

(c) Close Packing in Three Dimensions

All real structures are three dimensional structures. They can be obtained by stacking two dimensional layers one above the other. In the last Section, we discussed close packing in two dimensions which can be of two types; square close-packed and hexagonal close-packed. Let us see what types of three dimensional close packing can be obtained from these.

(i) ***Three dimensional close packing from two dimensional square close-packed layers:*** While placing the second square close-packed layer above the first we follow the same rule that was followed when one row was placed adjacent to the other. The second layer is placed over the first layer such that the spheres of the upper layer are exactly above those of the first layer. In this arrangement spheres of both the layers are perfectly aligned horizontally as well as vertically as shown in figure 1.15. Similarly, we may place more layers one above the other. If the arrangement of spheres in the first layer is called 'A' type, all the layers have the same arrangement. Thus this lattice has AAA.... type pattern. The lattice thus generated is the simple cubic lattice, and its unit cell is the primitive cubic unit cell .

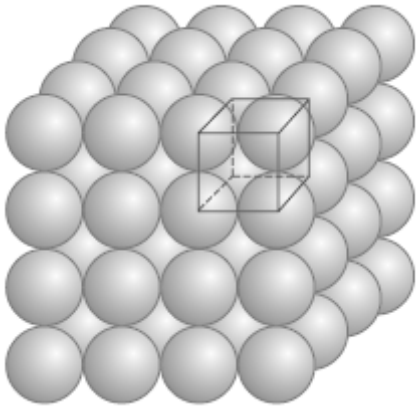


Fig.1.15 Simple cubic lattice formed by A A A… arrangement

(ii) ***Three dimensional close packing from two dimensional hexagonal close packed layers:*** Three dimensional close packed structure can be generated by placing layers one over the other.

(a) Placing second layer over the first layer Let us take a two dimensional hexagonal close packed layer 'A' and place a similar layer above it such that the spheres of the second layer are placed in the depressions of the first layer. Since the spheres of the two layers are aligned differently, let us call the second layer as B. It can be observed from figure 1.16 that not all the triangular voids of the first layer are covered by the spheres of the second layer. This gives rise to different arrangements. Wherever a sphere of the second layer is above the void of the first layer (or vice versa) a tetrahedral void is formed.

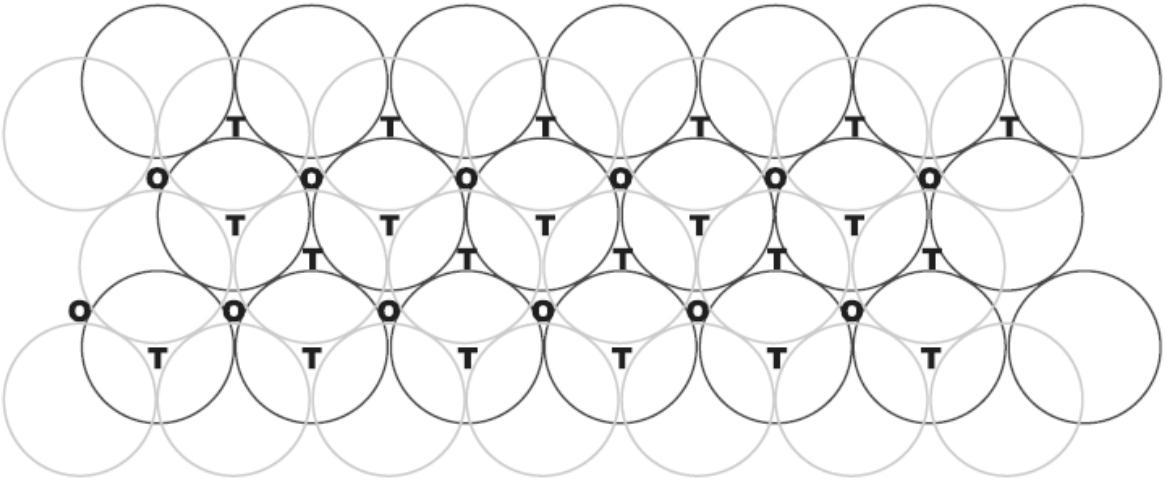


Fig.1.16 A stack of two layers of close packed spheres and voids generated in them. T = Tetrahedral void; O = Octahedral void

These voids are called **tetrahedral voids** because a ***tetrahedron*** is formed when the centres of these four spheres are joined. They have been marked as 'T' in figure 1.16. One such void has been shown separately in figure. 1.17.

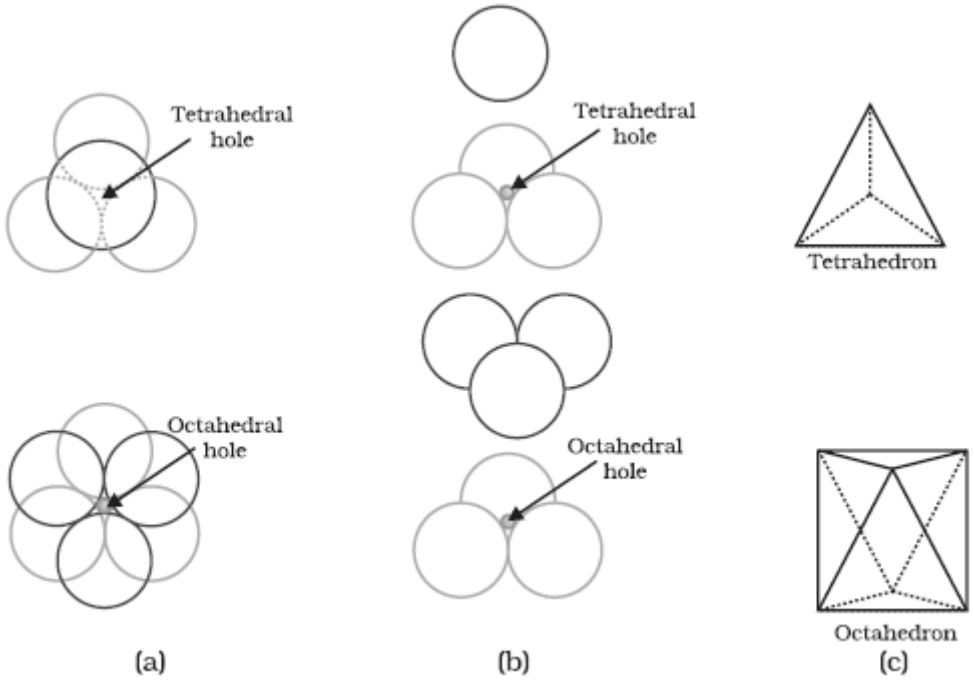


Fig.1.17 Tetrahedral and octahedral voids (a) top view (b) exploded side view and (c) geometrical shape of the void.

At other places, the triangular voids in the second layer are above the triangular voids in the first layer, and the triangular shapes of these do not overlap. One of them has the apex of the triangle pointing upwards and the other downwards. These voids have been marked as 'O' in figure 1.16. Such voids are surrounded by six spheres and are called **octahedral voids**. One such void has been shown separately in Fig. 1.17. The number of these two types of voids depends upon the number of close packed spheres.

Let the number of close packed spheres be *N*, then:

The number of octahedral voids generated = *N*

The number of tetrahedral voids generated = *2N*

(b) Placing third layer over the second layer When third layer is placed over the second, there are two possibilities.

(i) ***Covering Tetrahedral Voids*:** Tetrahedral voids of the second layer may be covered by the spheres of the third layer. In this case, the spheres of the third layer are exactly aligned with those of the first layer. Thus, the pattern of spheres is repeated in alternate layers. This pattern is often written as ABAB ..... pattern. This structure is called hexagonal close packed (*hcp*) structure. This sort of arrangement of atoms is found in many metals like magnesium and zinc.

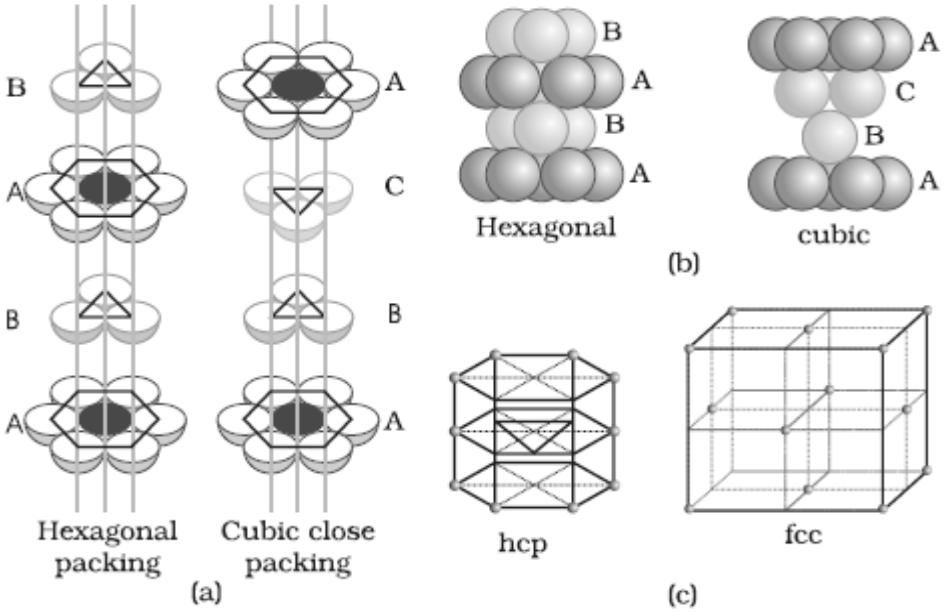


Fig.1.18 (a) Hexagonal cubic close-packing exploded view showing stacking of layers of spheres (b) four layers stacked in each case and (c) geometry of packing.

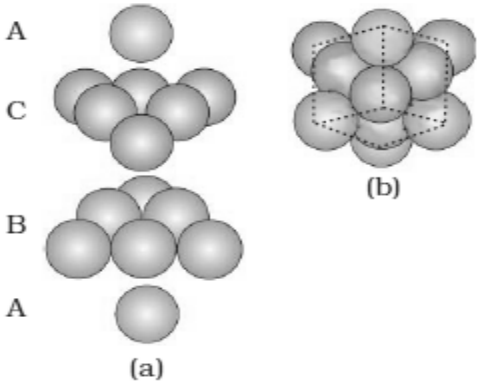


Fig. 1.19 (a) ABCABC... arrangement of layers when octahedral void is covered (b) fragment of structure formed by this arrangement resulting in cubic closed packed (ccp) or face centred cubic (fcc) structure.

(ii) ***Covering Octahedral Voids*:** The third layer may be placed above the second layer in a manner such that its spheres cover the octahedral voids. When placed in this manner, the spheres of the third layer are not aligned with those of either the first or the second layer. This arrangement is called "C' type. Only when fourth layer is placed, its spheres are aligned with those of the first layer as shown in figure 1.19. This pattern of layers is often written as ABCABC ..... This

structure is called cubic close packed (*ccp*) or face-centred cubic (*fcc*) structure. Metals such as copper and silver crystallise in this structure.

Both these types of close packing are highly efficient and 74% space in the crystal is filled. In either of them, each sphere is in contact with twelve spheres. Thus, the coordination number is 12 in either of these two structures.

1.6.2 Formula of a Compound and Number of Voids Filled

Earlier in the section, we have learnt that when particles are close packed resulting in either *ccp* or *hcp* structure, two types of voids are generated. While the number of octahedral voids present in a lattice is equal to the number of close packed particles, the number of tetrahedral voids generated is twice this number. In ionic solids, the bigger ions (usually anions) form the close packed structure and the smaller ions (usually cations) occupy the voids. If the latter ion is small enough then tetrahedral voids are occupied, if bigger, then octahedral voids. Not all octahedral or tetrahedral voids are occupied. In a given compound, the fraction of octahedral or tetrahedral voids that are occupied depends upon the chemical formula of the compound, as can be seen from the following examples.

Locating Tetrahedral and Octahedral Voids

We know that close packed structures have both tetrahedral and octahedral voids. Let us take *ccp* (or *fcc*) structure and locate these voids in it.

(a) Locating Tetrahedral Voids

Let us consider a unit cell of *ccp* or *fcc* lattice [Fig. 1(a)]. The unit cell is divided into eight small cubes.

Each small cube has atoms at alternate corners [figure 1.20(a)]. In all, each small cube has 4 atoms. When joined to each other, they make a regular tetrahedron. Thus, there is one tetrahedral void in each small cube and eight tetrahedral voids in total. Each of the eight small cubes have one void in one unit cell of *ccp* structure. We know that *ccp* structure has 4 atoms per unit cell. Thus, the number of tetrahedral voids is twice the number of atoms.

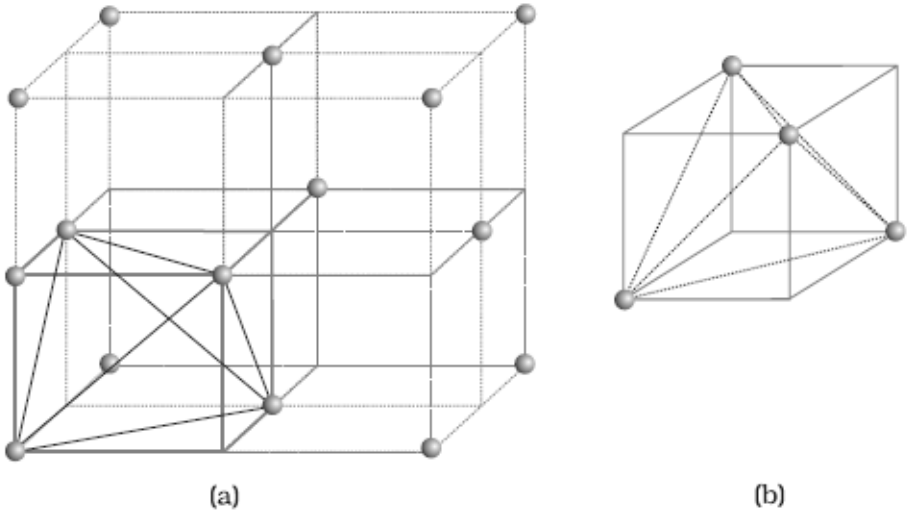


Fig1.20 (a) Eight tetrahedral voids per unit cell of ccp structure (b) one tetrahedral void showing the geometry.

(b) Locating Octahedral Voids

Let us again consider a unit cell of *ccp* or *fcc* lattice [Fig. 2(a)]. The body centre of the cube, C is not occupied but it is surrounded by six atoms on face centres. If these face centres are joined, an octahedron is generated. Thus, this unit cell has one octahedral void at the body centre of the cube.

Besides the body centre, there is one octahedral void at the centre of each of the 12 edges. [Fig. 2(b)]. It is surrounded by six atoms, three belonging to the same unit cell (2 on the corners and 1 on face centre) and three belonging to two adjacent unit cells. Since each edge of the cube is shared between four adjacent unit cells, so is the octahedral void located on it. Only $\frac{1}{4}^{th}$ of each void belongs to a particular unit cell.

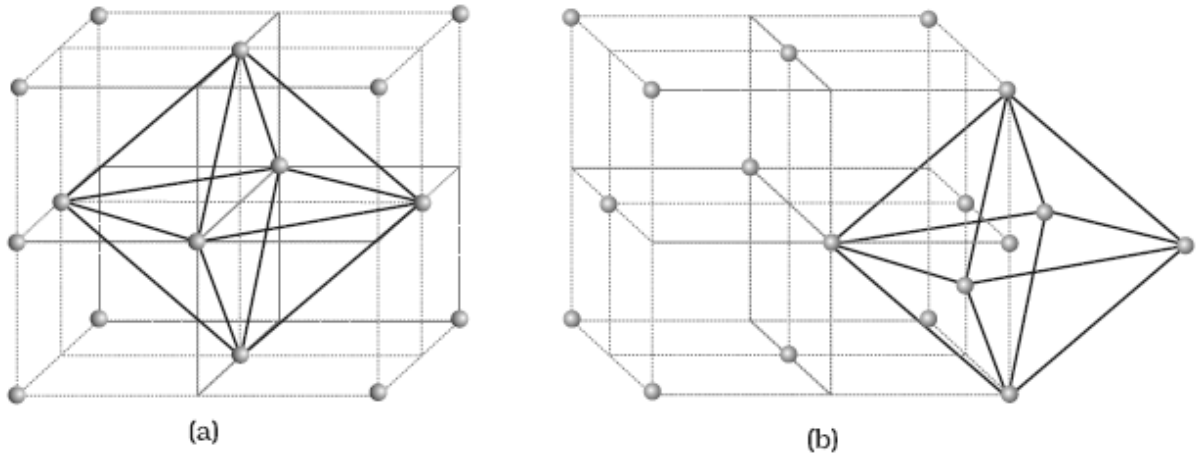


Fig 1.21 Location of octahedral voids per unit cell of ccp or fcc lattice (a) at the body centre of the cube and (b) at the centre of each edge (only one such void is shown).

Thus in *cubic close packed structure*:
Octahedral void at the body-centre of the cube = 1
12 octahedral voids located at each edge and shared between four unit cells.

$$= 12 \times \frac{1}{4} = 3$$

$\therefore$ Total number of octahedral voids = 4
We know that in ccp structure, each unit cell has 4 atoms. Thus, the number of octahedral voids is equal to this number.

1.7 Packing Efficiency

In whatever way the constituent particles (atoms, molecules or ions) are packed, there is always some free space in the form of voids. **Packing efficiency** is the percentage of total space filled by the particles. Let us calculate the packing efficiency in different types of structures.

1.7.1Packing Efficiency in hcp and ccp Structures

Both types of close packing (*hcp* and *ccp*) are equally efficient. Let us calculate the efficiency of packing in ccp structure. In Fig. 1.22 let the unit cell edge length be ‘a’ and face diagonal AC = b.

In Δ ABC

$AC^2 = b^2 = BC^2 + AB^2$

$= a^2 + a^2 = 2a^2$ or

$b = \sqrt{2}a$

If r is the radius of the sphere, we find

$b = 4r = \sqrt{2}a$

or $a = \frac{4r}{\sqrt{2}} = 2\sqrt{2}r$

(we can also write, $r = \frac{a}{2\sqrt{2}}$)

We know, that each unit cell in ccp structure, has effectively 4 spheres. Total volume of four spheres is equal to $4 \times (4/3)\pi r3$ and volume of the cube is a^3 or $\left(2\sqrt{2}r\right)^3$

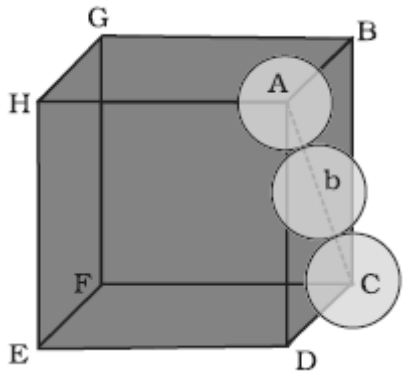


Fig.1.22 Cubic close packing other sides are not provided with spheres for sake of clarity.

Therefore,

$$\text{Packing efficiency} = \frac{\text{Volume occupied by four spheres in the unit cell} \times 100}{\text{Total volume of the unit cell}} \%$$

$$= \frac{4 \times (4/3)\pi r^3 \times 100}{\left(2\sqrt{2}r\right)^3} \%$$

$$= \frac{(16/3)\pi r^3 \times 100}{16\sqrt{2}r^3} \% = 74\%$$

1.7.2 Efficiency of Packing in Body-Centred Cubic Structsures

From below Figure 1.23, it is clear that the atom at the centre will be in touch with the other two atoms diagonally arranged.

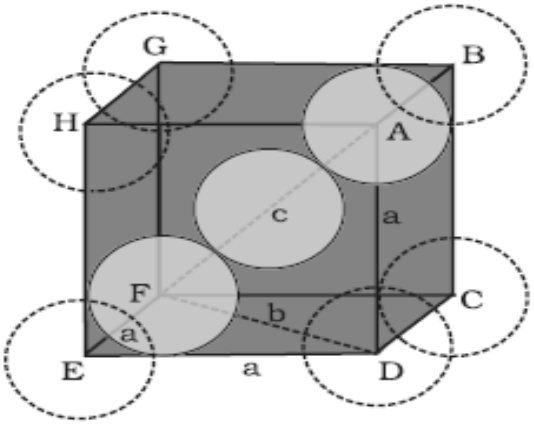


Fig.1.23 Body-centred cubic unit cell (sphere along the body diagonal are shown with solid boundaries).

In Δ EFD,

$$b^2 = a^2 + a^2 = 2a^2$$

$$b = \sqrt{2}a$$

Now in Δ AFD

$$c^2 = a^2 + b^2 = a^2 + 2a^2 = 3a^2$$

$$c = \sqrt{3}a$$

The length of the body diagonal c is equal to 4r, where r is the radius of the sphere (atom), as all the three spheres along the diagonal touch each other.

Therefore, $\sqrt{3}a = 4r$

$$a = \frac{4r}{\sqrt{3}}$$

Also we can write, $r = \frac{\sqrt{3}}{4}a$

In this type of structure, total number of atoms is 2 and their volume is $2 \times \left(\frac{4}{3}\right)\pi r^3$.

Volume of the cube, a^3 will be equal to $\left(\frac{4}{\sqrt{3}}r\right)^3$ or $a^3 = \left(\frac{4}{\sqrt{3}}r\right)^3$.

Therefore,

$$\text{Packing efficiency} = \frac{\text{Volume occupied by two spheres in the unit cell} \times 100}{\text{Total volume of the unit cell}}\%$$

$$= \frac{2 \times (4/3)\pi r^3 \times 100}{\left[(4/\sqrt{3})r\right]^3}\%$$

$$= \frac{(8/3)\pi r^3 \times 100}{64/(3\sqrt{3})r^3}\% = 68\%$$

1.7.3 Packing Efficiency in Simple Cubic Lattice

In a simple cubic lattice the atoms are located only on the corners of the cube. The particles touch each other along the edge (Fig. 1.24). Thus, the edge length or side of the cube 'a', and the radius of each particle, r are related as

$$a = 2r$$

The volume of the cubic unit cell = $a_3 = (2r)_3 = 8r_3$ Since a simple cubic unit cell contains only 1 atom The volume of the occupied space $= \frac{4}{3}\pi r^3$

$$\therefore \text{Packing efficiency} = \frac{\text{Volume of one atom}}{\text{Volume of cubic unit cell}} \times 100\%$$

$$= \frac{\frac{4}{3}\pi r^3}{8r^3} \times 100 = \frac{\pi}{6} \times 100$$

$$= 52.36\% = 52.4\ \%$$

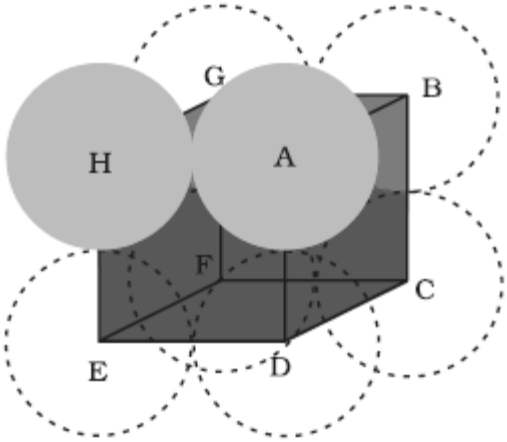


Fig.1.24 Simple cubic unit cell. The spheres are in contact with each other along the edge of the cube.

Thus, we may conclude that *ccp* and *hcp* structures have maximum packing efficiency.

Calculations Involving Unit Cell Dimensions

From the unit cell dimensions, it is possible to calculate the volume of the unit cell. Knowing the density of the metal, we can calculate the mass of the atoms in the unit cell. The determination of the mass of a single atom gives an accurate method of determination of Avogadro constant. Suppose, edge length of a unit cell of a cubic crystal determined by X-ray diffraction is *a, d* the density of the solid substance and *M* the molar mass. In case of cubic crystal:

Volume of a unit cell = *a3*

Mass of the unit cell

= number of atoms in unit cell × mass of each atom = $z \times m$

(Here *z* is the number of atoms present in one unit cell and *m* is the mass of a single atom)

Mass of an atom present in the unit cell:

$$m = \frac{M}{N_A} \text{ (} M \text{ is molar mass)}$$

Therefore, density of the unit cell

$$= \frac{mass\ of\ unit\ cell}{volume\ of\ unit\ cell}$$

$$= \frac{z.m}{a^3} = \frac{z.M}{a^3.N_A} \text{ or } d = \frac{zM}{a^3 N_A}$$

Remember, the density of the unit cell is the same as the density of the substance. The density of the solid can always be determined by other methods. Out of the five parameters (*d*, z *M*, *a* and *NA*), if any four are known, we can determine the fifth.

1.8 Imperfections in Solids

Although crystalline solids have short range as well as long range order in the arrangement of their constituent particles, yet crystals are not perfect. Usually a solid consists of an aggregate of large number of small crystals. These small crystals have defects in them. This happens when crystallisation process occurs at fast or moderate rate. Single crystals are formed when the process of crystallisation occurs at extremely slow rate. Even these crystals are not free of defects. The defects are basically irregularities in the arrangement of constituent particles. Broadly speaking, the defects are of two types, namely, *point defects* and *line defects*. **Point defects** are the irregularities or deviations from ideal arrangement around a point or an atom in a crystalline substance, whereas the *line defects* are the irregularities or deviations from ideal arrangement in entire rows of lattice points. These irregularities are called *crystal defects*. We shall confine our discussion to point defects only.

1.8.1 Types of Point Defects

Point defects can be classified into three types:

(i) stoichiometric defects
(ii) impurity defects and
(iii) non-stoichiometric defects.

(a) ***Stoichiometric Defects*** These are the point defects that do not disturb the stoichiometry of the solid. They are also called *intrinsic* or **thermodynamic defects.** Basically these are of two types, vacancy defects and interstitial defects.

(i) *Vacancy Defect*: When some of the lattice sites are vacant, the crystal is said to have **vacancy defect** (Fig. 1.25). This results in decrease in density of the substance. This defect can also develop when a substance is heated.

(ii) *Interstitial Defect*: When some constituent particles (atoms or molecules) occupy an interstitial site, the crystal is said to have **interstitial defect** (Fig. 1.26). This defect increases the density of the substance.

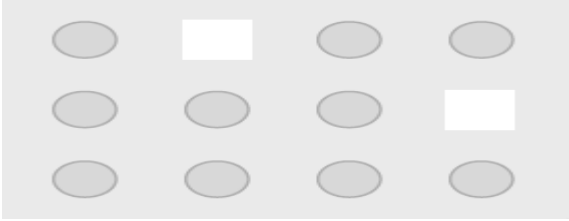


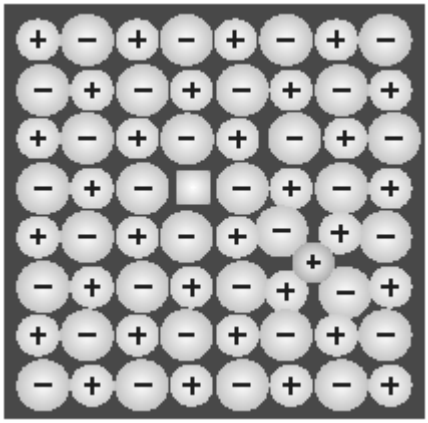
Fig.1.25 Vacancy defects

Vacancy and interstitial defects as explained above can be shown by non-ionic solids. Ionic solids must always maintain electrical neutrality. Rather than

simple vacancy or interstitial defects, they show these defects as **Frenkel and Schottky defects.**

(iii) *Frenkel Defect*: This defect is shown by ionic solids. The smaller ion (usually cation) is dislocated from its normal site to an interstitial site. It creates a *vacancy defect* at its original site and an **interstitial defect** at its new location. Frenkel defect is also called **dislocation defect**. It does not change the density of the solid. Frenkel defect is shown by ionic substance in which there is a large difference in the size of ions, for example, ZnS, AgCl, AgBr and AgI due to small size of Zn^{2+} and Ag^{+} ions.

(iv) *Schottky Defect*: It is basically a vacancy defect in ionic solids. In order to maintain electrical neutrality, the number of missing cations and anions are equal.



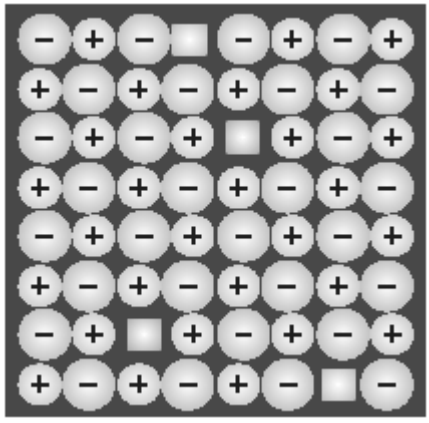


Fig. 1.26 Frenkel defects **Fig.1.27 Schottky defects**

Like simple vacancy defect, Schottky defect also decreases the density of the substance. Number of such defects in ionic solids is quite significant. For example, in NaCl there are approximately 10^6 Schottky pairs per cm^3 at room temperature. In 1 cm^3 there are about 10^{22} ions. Thus, there is one Schottky defect per 10^{16} ions. Schottky defect is shown by ionic substances in which the cation and anion are of almost similar sizes. For example, NaCl, KCl, CsCl and AgBr. It may be noted that AgBr shows both, Frenkel as well as Schottky defects.

(b) Impurity Defects

If molten NaCl containing a little amount of $SrCl_2$ is crystallised, some of the sites of Na+ ions are occupied by Sr^{2+} . Each Sr^{2+} replaces two Na^+ ions. It occupies the site of one ion and the other site remains vacant. The cationic vacancies thus produced are equal in number to that of Sr^{2+} ions. Another similar example is the solid solution of $CdCl_2$ and AgCl.

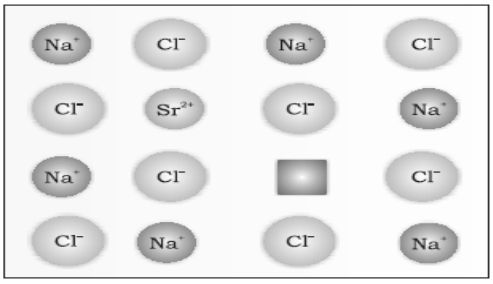


Fig.1.28 Introduction of cation vacancy in NaCl by substitution of Na^+ by Sr^{2+}

(b) ***Non-Stoichiometric Defects*** The defects discussed so far do not disturb the stoichiometry of the crystalline substance. However, a large number of nonstoichiometric inorganic solids are known which contain the constituent elements in non-stoichiometric ratio due to defects in their crystal structures. These defects are of two types: (i) metal excess defect and (ii) metal deficiency defect.

- *Metal Excess Defect : Metal excess defect due to anionic vacancies*: Alkali halides like NaCl and KCl show this type of defect. When crystals of NaCl are heated in an atmosphere of sodium vapour, the sodium atoms are deposited on the surface of the crystal. The Cl– ions diffuse to the surface of the crystal and combine with Na atoms to give NaCl. This happens by loss of electron by sodium atoms to form Na+ ions. The released electrons diffuse into the crystal and occupy anionic sites (Fig. 1.29). As a result the crystal now has an excess of sodium. The anionic sites occupied by unpaired electrons are called *F-centres* (from the German word *Farbenzenter* for colour centre). They impart yellow colour to the crystals of NaCl. The colour results by excitation of these electrons when they absorb energy from the visible light falling on the crystals. Similarly, excess of lithium makes LiCl crystals pink and excess of potassium makes KCl crystals violet (or lilac).

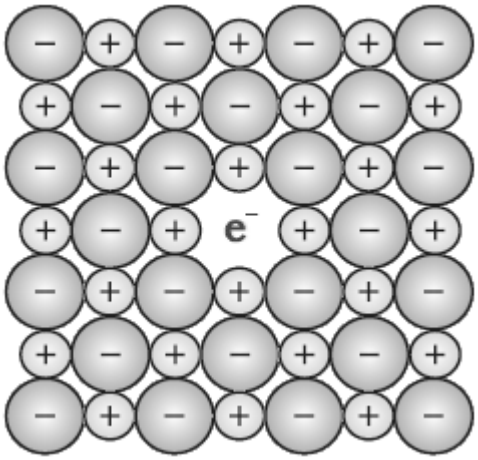


Fig.1.29 An F-centre in a crystal

- *Metal excess defect due to the presence of extra cations at interstitial sites*: Zinc oxide is white in colour at room temperature. On heating it loses oxygen and turns yellow.

$$ZnO \xrightarrow{\text{heating}} Zn^{2+} + \frac{1}{2}O_2 + 2e^-$$

Now there is excess of zinc in the crystal and its formula becomes $Zn_{1+x}O$. The excess Zn^{2+} ions move to interstitial sites and the electrons to neighbouring interstitial sites.

Metal Deficiency Defect

There are many solids which are difficult to prepare in the stoichiometric composition and contain less amount of the metal as compared to the stoichiometric proportion. A typical example of this type is FeO which is mostly found with a composition of $_{Fe0.95}O$. It may actually range from $Fe_{0.93}O$ to $Fe_{0.96}O$. In crystals of FeO some Fe^{2+} cations are missing and the loss of positive charge is made up by the presence of required number of Fe^{3+} ions.

1.9 Electrical Properties

Solids exhibit an amazing range of electrical conductivities, extending over 27 orders of magnitude ranging from 10^{-20} to 10^{7} ohm^{-1} m^{-1}. Solids can be classified into three types on the basis of their conductivities.

(i) *Conductors*: The solids with conductivities ranging between 10^{4} to 10^{7} $ohm^{-1}m^{-1}$ are called conductors. Metals have conductivities in the order of 10^{7} $ohm^{-1}m^{-1}$ are good conductors.

(ii) *Insulators:* These are the solids with very low conductivities ranging between 10^{-20} to 10^{-10} $ohm^{-1}m^{-1}$.

(iii) *Semiconductors:* These are the solids with conductivities in the intermediate range from 10^{-6} to 10^{4} $ohm^{-1}m^{-1}$.

1.9.1 Conduction of Electricity in Metals

A conductor may conduct electricity through movement of electrons or ions. Metallic conductors belong to the former category and electrolytes to the latter.

Metals conduct electricity in solid as well as molten state. The conductivity of metals depend upon the number of valence electrons available per atom. The atomic orbitals of metal atoms form molecular orbitals which are so close in energy to each other as to form a **band.** If this band is partially filled or it overlaps with a higher energy unoccupied conduction band, then electrons can flow easily under an applied electric field and the metal shows conductivity (Fig. 1.30 a).

If the gap between filled valence band and the next higher unoccupied band (conduction band) is large, electrons cannot jump to it and such a substance has very small conductivity and it behaves as an insulator (Fig. 1.30 b).

1.9.2 Conduction of Electricity in Semiconductors

In case of semiconductors, the gap between the valence band and conduction band is small. Therefore, some electrons may jump to conduction band and show some conductivity. Electrical conductivity of semiconductors increases with rise in temperature, since more electrons can jump to the conduction band. Substances like silicon and germanium show this type of behaviour and are called *intrinsic semiconductors.*

The conductivity of these intrinsic semiconductors is too low to be of practical use. Their conductivity is increased by adding an appropriate amount of suitable impurity. This process is called *doping*. Doping can be done with an impurity which is electron rich or electron deficient as compared to the intrinsic semiconductor silicon or germanium. Such impurities introduce *electronic effects* in them.

(a) Electron – rich impurities

Silicon and germanium belong to group 14 of the periodic table and have four valence electrons each. In their crystals each atom forms four covalent bonds with its neighbours. When doped with a group 15 element like P or As, which contains five valence electrons, they occupy some of the lattice sites in silicon or germanium crystal. Four out of five electrons are used in the formation of four covalent bonds with the four neighbouring silicon atoms. The fifth electron is extra and becomes delocalised. These delocalised electrons increase the conductivity of doped silicon (or germanium). Here the increase in conductivity is due to the *negatively* charged electron, hence silicon doped with electron-rich impurity is called *n*-type semiconductor.

(b) Electron – deficit impurities

Silicon or germanium can also be doped with a group 13 element like B, Al or Ga which contains only three valence electrons. The place where the fourth valence electron is missing is called *electron hole* or *electron vacancy*. An electron from a neighbouring atom can come and fill the electron hole, but in doing so it would leave an electron hole at its original position. If it happens, it would appear as if the electron hole has moved in the direction opposite to that of the electron that filled it. Under the influence of electric field, electrons would move towards the positively charged plate through electronic holes, but it would appear as if electron holes are positively charged and are moving towards negatively charged plate. This type of semi conductors are called *p-type* semiconductors.

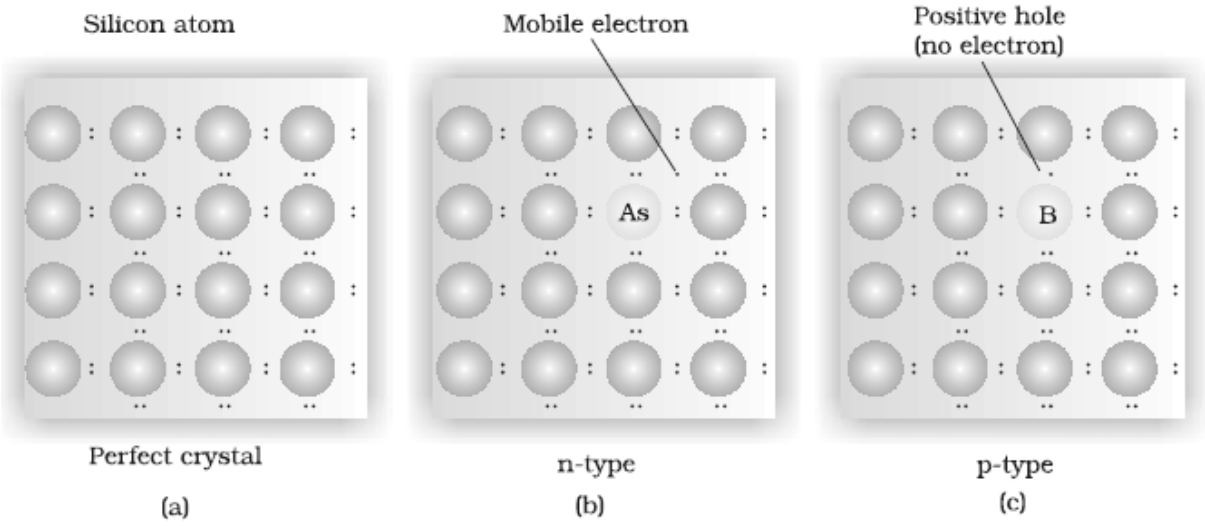


Fig.1.30 Creation of n-type and p-type semiconductors by doping groups 13 and 15 elements.

Applications of n-type and p-type semiconductors

Various combinations of *n*-type and *p*-type semiconductors are used for making electronic components. *Diode* is a combination of *n*-type and *p*-type semiconductors and is used as a rectifier. Transistors are made by sandwiching a layer of one type of semiconductor between two layers of the other type of semiconductor. *npn* and *pnp* type of transistors are used to detect or amplify radio or audio signals. The solar cell is an efficient photo-diode used for conversion of light energy into electrical energy.

Germanium and silicon are group 14 elements and therefore, have a characteristic valence of four and form four bonds as in diamond. A large variety of solid state materials have been prepared by combination of groups 13 and 15 or 12 and 16 to simulate average valence of four as in Ge or Si. Typical compounds of groups 13 – 15 are InSb, AlP and GaAs. Gallium arsenide (GaAs) semiconductors have very fast response and have revolutionised the design of semiconductor devices. ZnS, CdS, CdSe and HgTe are examples of groups 12 – 16 compounds. In these compounds, the bonds are not perfectly covalent and the ionic character depends on the electronegativities of the two elements.

It is interesting to learn that transition metal oxides show marked differences in electrical properties. TiO, CrO_2 and ReO_3 behave like metals. Rhenium oxide, ReO_3 is like metallic copper in its conductivity and appearance. Certain other oxides like VO, VO_2, VO_3 and TiO_3 show metallic or insulating properties depending on temperature.

1.7 Magnetic Properties

Every substance has some magnetic properties associated with it. The origin of these properties lies in the electrons. Each electron in an atom behaves like a tiny magnet. Its magnetic moment originates from two types of motions (i) its orbital motion around the nucleus and (ii) its spin around its own axis. Electron being a charged particle and undergoing these motions can be considered as a small loop of current which possesses a magnetic moment. Thus, each electron has a permanent spin and an orbital magnetic moment associated with it. Magnitude of this magnetic moment is very small and is measured in the unit called *Bohr magneton,* μ_B. It is equal to 9.27×10^{-24}A-m^2.

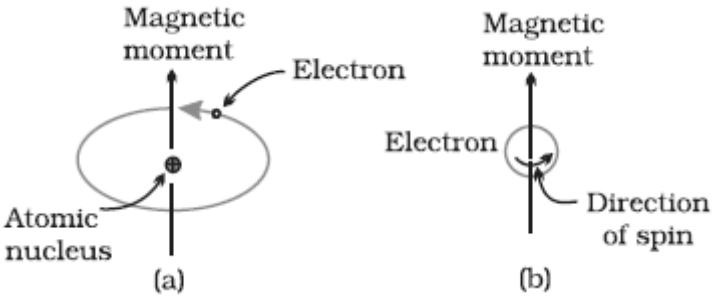


Fig.1.31 Demonstration of the magnetic moment associated with (a) an orbiting electron and (b) a spinning electron.

On the basis of their magnetic properties, substances can be classified into five categories: (i) paramagnetic (ii) diamagnetic (iii) ferromagnetic (iv) antiferromagnetic and (v) ferrimagnetic.

(i) *Paramagnetism*: Paramagnetic substances are weakly attracted by a magnetic field. They are magnetised in a magnetic field in the same direction. They lose their magnetism in the absence of magnetic field. Paramagnetism is due to presence of one or more unpaired electrons which are attracted by the magnetic field. O_2, Cu^{2+}, Fe^{3+}, Cr^{3+} are some examples of such substances.

(ii) *Diamagnetism*: Diamagnetic substances are weakly repelled by a magnetic field. H2O, NaCl and C6H6 are some examples of such substances. They are weakly magnetised in a magnetic field in opposite direction. Diamagnetism is shown by those substances in which all the electrons are paired and there are no unpaired electrons. Pairing of electrons cancels their magnetic moments and they lose their magnetic character.

(iii) *Ferromagnetism*: A few substances like iron, cobalt, nickel, gadolinium and CrO_2 are attracted very strongly by a magnetic field. Such substances are called ferromagnetic substances. Besides strong attractions, these substances can be permanently magnetised. In solid state, the metal ions of ferromagnetic substances are grouped together into small regions called *domains*. Thus, each domain acts as a tiny magnet. In an unmagnetised piece of a ferromagnetic substance the domains are randomly oriented and their magnetic moments get cancelled. When the substance is placed in a magnetic field all the domains get oriented in the direction of the magnetic field and a strong magnetic effect is produced. This ordering of domains persist even when the magnetic field is removed and the ferromagnetic substance becomes a permanent magnet.

(iv) *Antiferromagnetism*: Substances like MnO showing antiferromagnetism have domain structure similar to ferromagnetic substance, but their domains are oppositely oriented and cancel out each other's magnetic moment.

(v) *Ferrimagnetism*: Ferrimagnetism is observed when the magnetic moments of the domains in the substance are aligned in parallel and anti-parallel directions in unequal numbers. They are weakly attracted by magnetic field as compared to ferromagnetic substances. Fe_3O_4 (magnetite) and ferrites like $MgFe_2O_4$ and $ZnFe_2O_4$ are examples of such substances. These substances also lose ferrimagnetism on heating and become paramagnetic.

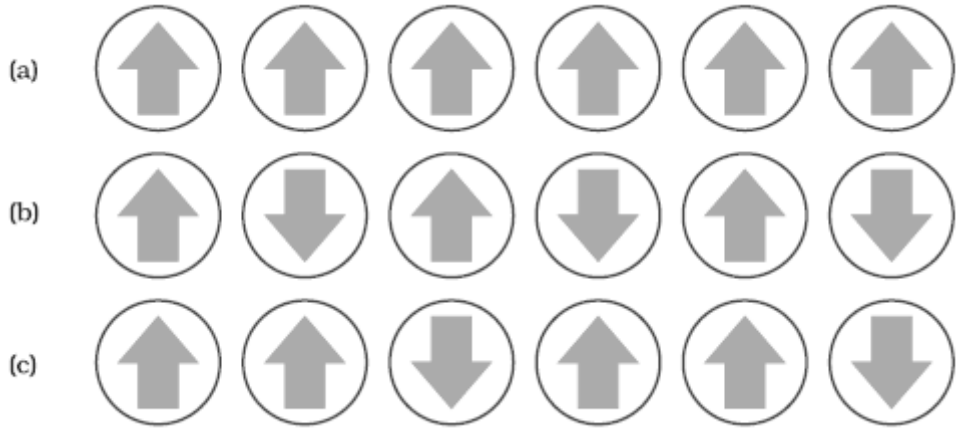


Fig.1.32 Schematic alignment of magnetic moments in (a) ferromagnetic (b) antiferromagnetic and (c) ferrimagnetic.

Summary

Solids have definite mass, volume and shape. This is due to the fixed position of their constituent particles, short distances and strong interactions between them. In **amorphous** solids, the arrangement of constituent particles has only **short range order** and consequently they behave like **super cooled liquids**, do not have sharp melting points and are isotropic in nature. In crystalline solids there is long range order in the arrangement of their constituent particles. They have sharp melting points, are anisotropic in nature and their particles have characteristic shapes. Properties of **crystalline** solids depend upon the nature of interactions between their constituent particles. On this basis, they can be divided into four categories, namely: **molecular**, **ionic**, **metallic** and **covalent** solids. They differ widely in their properties.

The constituent particles in crystalline solids are arranged in a regular pattern which extends throughout the crystal. This arrangement is often depicted in the form of a three dimensional array of points which is called crystal lattice. Each **lattice point** gives the location of one particle in space. In all, fourteen different types of lattices are possible which are called **Bravais lattices**. Each lattice can be generated by repeating its small characteristic portion called **unit cell**. A unit cell is characterised by its edge lengths and three angles between these edges. Unit cells can be either **primitive** which have particles only at their corner positions or **centred**. The centred unit cells have additional particles at their body centre (**bodycentred**), at the centre of each face (**face-centred**) or at the centre of two opposite faces (**end-centred**). There are seven types of **primitive unit** cells. Taking centred unit cells also into account, there are fourteen types of unit cells in all, which result in fourteen **Bravais lattices**.

Close-packing of particles result in two highly efficient lattices, **hexagonal close-packed** (**hcp**) and **cubic close-packed** (**ccp**). The latter is also called facecentred cubic (**fcc**) lattice. In both of these packings 74% space is filled. The remaining space is present in the form of two types of voids-octahedral voids and tetrahedral voids. Other types of packing are not close-packings and have less efficient packing of particles. While in **body-centred cubic lattice** (**bcc**) 68% space is filled, in simple cubic lattice only 52.4 % space is filled.

Solids are not perfect in structure. There are different types of **imperfections** or **defects** in them. Point defects and line defects are common types of defects. Point defects are of three types - **stoichiometric defects**, **impurity defects** and **non-stoichiometric defects**. **Vacancy defects** and **interstitial defects** are the two basic types of stoichiometric point

defects. In ionic solids, these defects are present as **Frenkel** and **Schottky defects**. Impurity defects are caused by the presence of an impurity in the crystal. In ionic solids, when the ionic impurity has a different valence than the main compound, some vacancies are created. Nonstoichiometric defects are of metal excess type and metal deficient type. Sometimes calculated amounts of impurities are introduced by **doping in semiconductors** that change their electrical properties. Such materials are widely used in electronics industry. Solids show many types of magnetic properties like **paramagnetism**, **diamagnetism**, **ferromagnetism**, **antiferromagnetism** and **ferrimagnetism**. These properties are used in audio, video and other recording devices. All these properties can be correlated with their electronic configurations or structures.

IMPORTANT QUESTIONS:

1. What is schootky defect
2. What is Frenkel defect
3. Derive Bragg's equation.
4. What are semiconductors
5. Define the following terms
(a) Tyndall effect (b) Brownian movement (c) Electrophoresis
(d) Coagulation.

CHAPTER 2

SOLUTIONS

In this Unit, we will consider mostly liquid solutions and their formation. This will be followed by studying the properties of the solutions, like vapour pressure and colligative properties. We will begin with types of solutions and then various alternatives in which concentrations of a solute can be expressed in liquid solution.

2.1 Types of Solution

Solutions are **homogeneous** mixtures of two or more than two components. By homogenous mixture we mean that its composition and properties are uniform throughout the mixture. Generally, the component that is present in the largest quantity is known as **solvent**. Solvent determines the physical state in which solution exists. One or more components present in the solution other than solvent are called **solutes.** In this Unit we shall consider only **binary solutions** (i.e., consisting of two components). Here each component may be solid, liquid or in gaseous state and are summarised in Table 2.1.

Table 2.1. Types of Solutions

Type of Solution	Solute	Solvent	Common Examples
Gaseous Solutions	Gas	Gas	Mixture of oxygen and nitrogen gases
	Liquid	Gas	Chloroform mixed with nitrogen gas
	Solid	Gas	Camphor in nitrogen gas
Liquid Solutions	Gas	Liquid	Oxygen dissolved in water
	Liquid	Liquid	Ethanol dissolved in water
	Solid	Liquid	Glucose dissolved in water
Solid Solutions	Gas	Solid	Solution of hydrogen in palladium
	Liquid	Solid	Amalgam of mercury with sodium
	Solid	Solid	Copper dissolved in gold

2.2 Expressing Concentration of Solutions

Composition of a solution can be described by expressing its concentration. The latter can be expressed either qualitatively or quantitatively. For example, qualitatively we can say that the solution is dilute (i.e., relatively very small quantity of solute) or it is concentrated (i.e., relatively very large quantity of solute). But in real life these kinds of description can add to lot of confusion and thus the need for a quantitative description of the solution.

There are several ways by which we can describe the concentration of the solution quantitatively.

(i) Mass percentage (w/w): The mass percentage of a component of a solution is defined as:

$$\text{Mass \% of a component} = \frac{\text{Mass of the component in the solution}}{\text{Total mass of the solution}} \times 100 \qquad (2.1)$$

For example, if a solution is described by 10% glucose in water by mass, it means that 10 g of glucose is dissolved in 90 g of water resulting in a 100 g solution. Concentration described by mass percentage is commonly used in industrial chemical applications. For example, commercial bleaching solution contains 3.62 mass percentage of sodium hypochlorite in water.

(ii) Volume percentage (v/v): The volume percentage is defined as:

$$\text{Volume of the Component} = \frac{\text{Volume of the component}}{\text{Total volume of solution}} \text{ X } 100 \quad (2.2)$$

For example, 10% ethanol solution in water means that 10 mL of ethanol is dissolved in water such that the total volume of the solution is 100 mL. Solutions containing liquids are commonly expressed in this unit. For example, a 35% (*v*/*v*) solution of ethylene glycol, an antifreeze, is used in cars for cooling the engine. At this concentration the antifreeze lowers the freezing point of water to 255.4K (–17.6°C).

(iii) Mass by volume percentage (w/v): Another unit which is commonly used in medicine and pharmacy is mass by volume percentage. It is the mass of solute dissolved in 100 mL of the solution.

(iv) Parts per million: When a solute is present in **trace** quantities, it is convenient to express concentration in **parts per million (ppm)** and is defined as:

$$\text{Parts per million} = \frac{\text{Number of parts of the component}}{\text{Total number of parts of all components of the solution}} \times 10^6 \quad (2.3)$$

As in the case of percentage, concentration in parts per million can also be expressed as mass to mass, volume to volume and mass to volume. A litre of sea water (which weighs 1030 g) contains about 6×10^{-3} g of dissolved oxygen (O_2). Such a small concentration is also expressed as 5.8 g per 10^6 g (5.8 ppm) of sea water. The concentration of pollutants in water or atmosphere is often expressed in terms of μg mL^{-1} or ppm.

(v) Mole fraction: Commonly used symbol for mole fraction is *x* and subscript used on the right hand side of *x* denotes the component. It is defined as:

$$\text{Mole fraction of a component} = \frac{\text{Number of moles of the component}}{\text{Total number of moles of all the components}} \quad (2.4)$$

For example, in a binary mixture, if the number of moles of A and B are n_A and n_B respectively, the mole fraction of A will be

$$x_A = \frac{n_A}{n_A + n_B} \quad (2.5)$$

For a solution containing i number of components, we have:

$$x_i = \frac{n_i}{n_1 + n_2 + \cdots + n_i} = \frac{n_i}{\sum n_i} \quad (2.6)$$

It can be shown that in a given solution sum of all the mole fractions is unity, i.e.

$$x_1 + x_2 + \ldots\ldots\ldots\ldots\ldots + x_i = 1 \quad (2.7)$$

Mole fraction unit is very useful in relating some physical properties of solutions, say vapour pressure with the concentration of the solution and quite useful in describing the calculations involving gas mixtures.

(vi) ***Molarity***: Molarity (*M*) is defined as number of moles of solute dissolved in one litre (or one cubic decimetre) of solution,

$$\text{Molarity} = \frac{\text{Moles of solute}}{\text{Volume of solution in litre}} \quad (2.8)$$

For example, 0.25 mol L^{-1} (or 0.25 M) solution of NaOH means that 0.25 mol of NaOH has been dissolved in one litre (or one cubic decimeter).

(vii) ***Molality*:** Molality (*m*) is defined as the number of moles of the solute per kilogram (kg) of the solvent and is expressed as:

$$\text{Molality (m)} = \frac{\text{Moles of solute}}{\text{Mass of solvent in kg}} \quad (2.9)$$

For example, 1.00 mol kg^{-1} (or 1.00 m) solution of KCl means that 1 mol (74.5 g) of KCl is dissolved in 1 kg of water.

Each method of expressing concentration of the solutions has its own merits and demerits. Mass %, ppm, mole fraction and molality are independent of temperature, whereas molarity is a function of temperature. This is because volume depends on temperature and the mass does not.

2.3 Solubility

Solubility of a substance is its maximum amount that can be dissolved in a specified amount of solvent. It depends upon the nature of solute and solvent as well as temperature and pressure. Let us consider the effect of these factors in solution of a solid or a gas in a liquid.

2.3.1 Solubility of a Solid in a Liquid

Every solid does not dissolve in a given liquid. While sodium chloride and sugar dissolve readily in water, naphthalene and anthracene do not. On the other hand, naphthalene and anthracene dissolve readily in benzene but sodium chloride and sugar do not. It is observed that polar solutes dissolve in polar solvents and non polar solutes in non-polar solvents. In general, a solute dissolves in a solvent if the intermolecular interactions are similar in the two or we may say **like dissolves like**.

When a solid solute is added to the solvent, some solute dissolves and its concentration increases in solution. This process is known as dissolution. Some solute particles in solution collide with the solid solute particles and get separated out of solution. This process is known as crystallisation**.** A stage is reached when the two processes occur at the same rate. Under such conditions, number of solute particles going into solution will be equal to the solute particles separating out and a state of dynamic equilibrium is reached.

$$solute + solvent \rightleftharpoons solution \quad 2.10$$

At this stage the concentration of solute in solution will remain constant under the given conditions, i.e., temperature and pressure. Similar process is followed when gases are dissolved in liquid solvents. Such a solution in which no more solute can be dissolved at the same temperature and pressure is called a **saturated solution**. An *unsaturated solution* is one in which more solute can be dissolved at the same temperature. The solution which is in dynamic equilibrium with undissolved solute is the saturated solution and contains the maximum amount of solute dissolved in a given amount of solvent. Thus, the concentration of solute in such a solution is its solubility.

Earlier we have observed that solubility of one substance into another depends on the nature of the substances. In addition to these variables, two other parameters, i.e., temperature and pressure also control this phenomenon.

Effect of temperature

The solubility of a solid in a liquid is significantly affected by temperature changes. Consider the equilibrium represented by equation 2.10. This, being dynamic equilibrium, must follow **Le Chateliers Principle**. In general, if in a *nearly saturated solution*, the dissolution process is endothermic ($_{sol}H > 0$), the solubility should increase with rise in temperature and if it is exothermic ($_{sol}H > 0$) the solubility should decrease. These trends are also observed experimentally.

Effect of pressure

Pressure does not have any significant effect on solubility of solids in liquids. It is so because solids and liquids are highly incompressible and practically remain unaffected by changes in pressure.

2.3.2 Solubility of a Gas in a Liquid

Many gases dissolve in water. Oxygen dissolves only to a small extent in water. It is this dissolved oxygen which sustains all aquatic life. On the other hand, hydrogen chloride gas (HCl) is highly soluble in water. Solubility of gases in liquids is greatly affected by pressure and temperature. The solubility of gases is increase with increase of pressure. For solution of gases in a solvent, consider a system as shown in Fig. 2.1 (a). The lower part is solution and the upper part is gaseous system at pressure *p* and temperature *T*. Assume this system to be in a state of dynamic equilibrium, i.e., under these conditions rate of gaseous particles entering and leaving the solution phase is the same. Now increase the pressure over the solution phase by compressing the gas to a smaller volume [Fig. 2.1 (b)]. This will increase the number of gaseous particles per unit volume over the solution and also the rate at which the gaseous particles are striking the surface of solution to enter it. The solubility of the gas will increase until a new equilibrium is reached resulting in an increase in the pressure of a gas above the solution and thus its solubility increases.

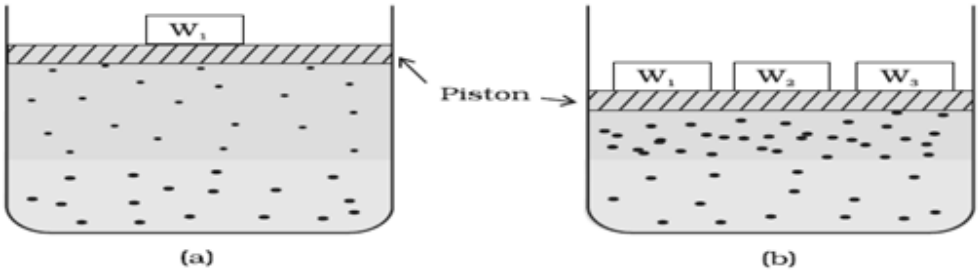


Fig. 2.1 Effect of pressure on the solubility of a gas. The concentration of dissolved gas is proportional to the pressure on the gas above the solution.

Henry was the first to give a quantitative relation between pressure and solubility of a gas in a solvent which is known as **Henry's law**. The law states that at a constant temperature**, the solubility of a gas in a liquid is directly proportional to the pressure of the gas**. Dalton, a contemporary of Henry, also concluded independently that the solubility of a gas in a liquid solution is a function of partial pressure of the gas. If we use the mole fraction of a gas in the solution as a measure of its solubility, then it can be said that the **mole fraction of gas in the**

solution is proportional to the partial pressure of the gas over the solution. The most commonly used form of Henry's law states that "**the partial pressure of the gas in vapour phase (*p*) is proportional to the mole fraction of the gas (*x*) in the solution"** and is expressed as:

$$p = K_H \; x \qquad (2.11)$$

Here *K*H is the Henry's law constant. If we draw a graph between partial pressure of the gas versus mole fraction of the gas in solution, then we should get a plot of the type as shown in Fig. 2.2.

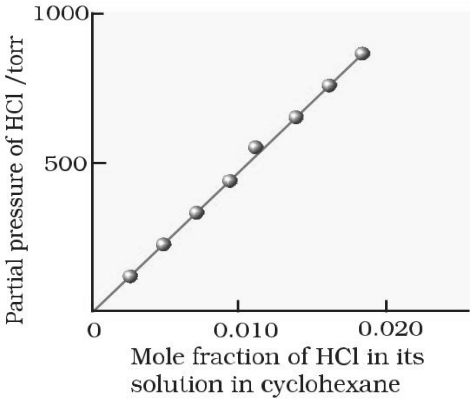


Fig. 2.2 Experimental results for the solubility of HCl gas in cyclohexane at 293 K. The slope of the line is the Henry's Law constant, K_H.

Different gases have different K_H values at the same temperature (Table 2.2). This suggests that K_H is a function of the nature of the gas.

It is obvious from equation (2.11) that higher the value of K_H at a given pressure, the lower is the solubility of the gas in the liquid. It can be seen from Table 2.2 that K_H values for both N_2 and O_2 increase with increase of temperature indicating that the solubility of gases increases with decrease of temperature. It is due to this reason that aquatic species are more comfortable in cold waters rather than in warm waters.

Henry's law finds several applications in industry and explains some biological phenomena. Notable among these are:

- To increase the solubility of CO_2 in soft drinks and soda water, the bottle is sealed under high pressure.

Table 2.2: Values of Henry's Law Constant for Some Selected Gases in Water

Gas	Temperature/K	K_H /kbar	Gas	Temperature/K	K_H/kbar
He	293	144.97	Argon	298	40.3
H_2	293	69.16	CO_2	298	1.67
N_2	293	76.48	Formaldehyde	298	1.83×10^{-5}
N_2	303	88.84	Methane	298	0.413
O_2	293	34.86			
$^{0}2$	303	46.82	Vinyl chloride	298	0.611

- Scuba divers must cope with high concentrations of dissolved gases while breathing air at high pressure underwater. Increased pressure increases the solubility of atmospheric gases in blood. When the divers come towards surface, the pressure gradually decreases. This releases the dissolved gases and leads to the formation of bubbles of nitrogen in the blood. This blocks capillaries and creates a medical condition known as *bends*, which are painful and dangerous to life. To avoid bends, as well as, the toxic effects of high concentrations of nitrogen in the blood, the tanks used by scuba divers are filled with air diluted with helium (11.7% helium, 56.2% nitrogen and 32.1% oxygen).
- At high altitudes the partial pressure of oxygen is less than that at the ground level. This leads to low concentrations of oxygen in the blood and tissues of people living at high altitudes or climbers. Low blood oxygen causes climbers to become weak and unable to think clearly, symptoms of a condition known as *anoxia*.

Effect of Temperature

Solubility of gases in liquids decreases with rise in temperature. When dissolved, the gas molecules are present in liquid phase and the process of dissolution can be considered similar to condensation and heat is evolved in this process. We have learnt in the last Section that dissolution process involves dynamic equilibrium and thus must follow Le Chatelier's Principle. As dissolution is an exothermic process, the solubility should decrease with increase of temperature.

2.4 Vapour Pressure of Liquid Solutions

Liquid solutions are formed when solvent is a liquid. The solute can be a gas, a liquid or a solid. Solutions of gases in liquids have already been discussed in Section 2.3.2. In this Section, we shall discuss the solutions of liquids and solids in a liquid. Such solutions may contain one or more volatile components. Generally, the liquid solvent is volatile. The solute may or may not be volatile. We shall discuss the properties of only binary solutions, that is, the solutions containing two components, namely, the solutions of (i) liquids in liquids and (ii) solids in liquids.

2.4.1 Vapour Pressure of Liquid- Liquid Solutions

Let us consider a binary solution of two volatile liquids and denote the two components as 1 and 2. When taken in a closed vessel, both the components would evaporate and eventually an equilibrium would be established between vapour phase and the liquid phase. Let the total vapour pressure at this stage be p_{total} and p_1 and p_2 be the partial vapour pressures of the two components 1 and 2 respectively. These partial pressures are related to the mole fractions x_1 and x_2 of the two components 1 and 2 respectively.

The French chemist, Francois Marte Raoult (1886) gave the quantitative relationship between them. The relationship is known as the **Raoult's law** which states that **for a solution of volatile liquids, the partial vapour pressure of each component in the solution is directly proportional to its mole fraction**.

Thus, for component 1

$$p_1 \propto x_1$$

and $$p_1 = p_1^0 x_1 \quad (2.12)$$

where p_1^0 is the vapour pressure of pure component 1 at the same temperature.

Similarly, for component 2

$$p_2 = p_2^0 x_2 \quad (2.13)$$

where p_2^0 represents the vapour pressure of the pure component 2. According to **Dalton's law of partial pressures**, the total pressure (p_{total}) over the solution phase in the container will be the sum of the partial pressures of the components of the solution and is given as:

$$p_{total} = p_1 + p_2 \quad (2.14)$$

(i) Substituting the values of p_1 and p_2, we get

$$p_{total} = x_1 p_1^0 + x_2 p_2^0$$
$$= (1 - x_2) p_1^0 + x_2 p_2^0 \quad (2.15)$$
$$= p_1^0 + x_2 (p_2^0 - p_1^0) \quad (2.16)$$

Following conclusions can be drawn from equation (2.16).

(ii) Total vapour pressure over the solution can be related to the mole fraction of any one component.

(iii) Total vapour pressure over the solution varies linearly with the mole fraction of component 2.

(iv) Depending on the vapour pressures of the pure components 1 and 2, total vapour pressure over the solution decreases or increases with the increase of the mole fraction of component 1

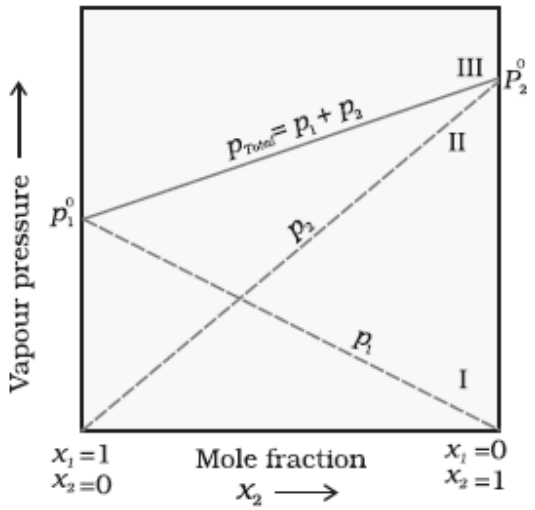


Fig. 2.3 The plot of vapour pressure and mole fraction of an ideal solution at constant temperature. The dashed lines I and II represent the partial pressure of the components. (It can be seen from the plot that p_1 and p_2 are directly proportional to x_1 and x_2, respectively). The total vapour pressure is given by line marked III in the figure.

A plot of p_1 or p_2 versus the mole fractions x_1 and x_2 for a solution gives a linear plot as shown in Fig. 2.3. These lines (I and II) pass through the points and respectively when x_1 and x_2 equal unity. Similarly the plot (line III) of $p$total versus x_2 is also linear (Fig. 2.3). The minimum value of $p$total is p_1^0 and the maximum value isp_2^0, assuming that component 1 is less volatile than component 2, i.e.,$p_1^0 < p_2^0$. The composition of vapour phase in equilibrium with the solution is determined by the partial pressures of the components. If y_1 and y_2 are the mole fractions of the components 1 and 2 respectively in the vapour phase then, using Dalton's law of partial pressures:

$$p_1 = y_1\, p_{\text{total}} \qquad (2.17)$$

$$p_2 = y_2\, p_{\text{total}} \qquad (2.18)$$

In general

$$p_i = y_i\ \ p_{\text{total}} \qquad (2.19)$$

2.4.2 Raoult's Law as a special case of Henry's Law

According to Raoult's law, the vapour pressure of a volatile component in a given solution is given by $p_i = x_i\, p_i^0$. In the solution of a gas in a liquid, one of the components is so volatile that it exists as a gas and we have already seen that its solubility is given by Henry's law which states that

$$p = K_H x.$$

If we compare the equations for Raoult's law and Henry's law, it can be seen that the partial pressure of the volatile component or gas is directly proportional to its mole fraction in solution. Only the proportionality constant K_H differs from p_1^0. Thus, Raoult's law becomes a special case of Henry's law in which K_H becomes equal to p_1^0.

2.4.3 Vapour Pressure of Solutions of Solids in Liquids

Another important class of solutions consists of solids dissolved in liquid, for example, sodium chloride, glucose, urea and cane sugar in water and iodine and sulphur dissolved in carbon disulphide. Some physical properties of these solutions are quite different from those of pure solvents. For example, vapour pressure. We have learnt in Unit 5, Class XI, that liquids at a given temperature vapourise and under equilibrium conditions the pressure exerted by the vapours of the liquid over the liquid phase is called vapour pressure [Fig. 2.4 (a)]. In a pure liquid the entire surface is occupied by the molecules of the liquid. If a non-volatile solute is added to a solvent to give a solution [Fig. 2.4.(b)], the vapour pressure of the solution is solely from the solvent alone. This vapour pressure of the solution at a given temperature is found to be lower than the vapour pressure of the pure solvent at the same temperature. In the solution, the surface has both solute and solvent molecules; thereby the fraction of the surface covered by the solvent molecules gets reduced. Consequently, the number of solvent molecules escaping from the surface is correspondingly reduced, thus, the vapour pressure is also reduced.

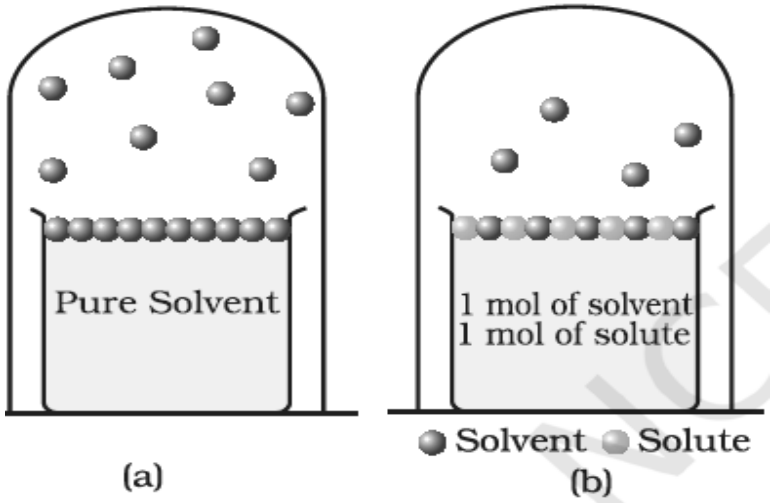


Fig. 2.4 Decrease in the vapour pressure of the solvent on account of the presence of solute in the solvent (a) evaporation of the molecules of the solvent from its surface is denoted by , (b) in a solution, solute particles have been denoted by and they also occupy part of the surface area.

The decrease in the vapour pressure of solvent depends on the quantity of non-volatile solute present in the solution, irrespective of its nature. For example, decrease in the vapour pressure of water by adding 1.0 mol of sucrose to one kg of water is nearly similar to that produced by adding 1.0 mol of urea to the same quantity of water at the same temperature. Raoult's law in its general form can be stated as, **for any solution the partial vapour pressure of each volatile component in the solution is directly proportional to its mole fraction**. In a binary solution, let us denote the solvent by 1 and solute by 2. When the solute is non-volatile, only the solvent molecules arepresent in vapour phase and contribute to vapour pressure. Let p_1 be the vapour pressure of the solvent, x_1 be its mole fraction, p_{i0} be its vapour pressure in the pure state. Then according to Raoult's law p_1 x_1 and $p_1 = x_1\ 0\ 1\ p$ (2.20) The proportionality constant is equal to the vapour pressure of pure solvent, $_{0\,1}p$. A plot between the vapour pressure and the mole fraction of the solvent is linear (Fig. 2.5).

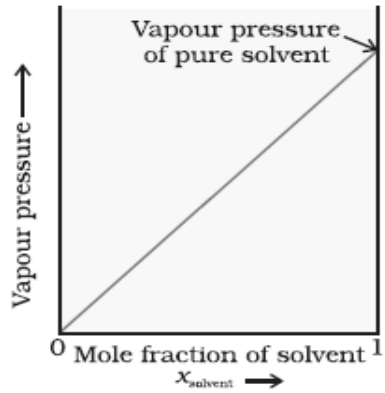


Fig. 2.5 If a solution obeys Raoult's law for all concentrations, its vapour pressure would vary linearly from zero to the vapour pressure of the pure solvent.

2.5 Ideal and Non-ideal Solutions

Liquid-liquid solutions can be classified into ideal and non-ideal solutions on the basis of Raoult's law.

2.5.1 Ideal Solutions

The solutions which obey Raoult's law over the entire range of concentration are known as *ideal solutions*. The ideal solutions have two other important properties. The enthalpy of mixing of the pure components to form the solution is zero and the volume of mixing is also zero, i.e.,

$$\Delta \text{mix}H = 0, \Delta \text{mix}V = 0 \qquad (2.21)$$

It means that no heat is absorbed or evolved when the components are mixed. Also, the volume of solution would be equal to the sum of volumes of the two components. At molecular level, ideal behaviour of the solutions can be explained by considering two components A and B. In pure components, the intermolecular attractive interactions will be of types A-A and B-B, whereas in the binary solutions in addition to these two interactions, A-B type of interactions will also be present. If the intermolecular attractive forces between the A-A and B-B are nearly equal to those between A-B, this leads to the formation of ideal solution. A perfectly ideal solution is rare but some solutions are nearly ideal in behaviour. Solution of n-hexane and n-heptane, bromoethane and chloroethane, benzene and toluene, etc. fall into this category.

2.5.2 Non-ideal Solutions

When a solution does not obey Raoult's law over the entire range of concentration, then it is called *non-ideal solution*. The vapour pressure of such a solution is either higher or lower than that predicted by Raoult's law (equation 2.16). If it is higher, the solution exhibits **positive deviation** and if it is lower, it exhibits **negative deviation** from Raoult's law. The plots of vapour pressure as a function of mole fractions for such solutions are shown in Fig. 2.6.

The cause for these deviations lie in the nature of interactions at the molecular level. In case of positive deviation from Raoult's law, A-B interactions are weaker than those between A-A or B-B, i.e., in this case the intermolecular attractive forces between the solute-solvent molecules are weaker than those between the solute-solute and solvent-solvent molecules. This means that in such solutions, molecules of A (or B) will find it easier to escape than in pure state. This will increase the vapour pressure and result in positive deviation. Mixtures of ethanol and acetone behave in this manner. In pure ethanol, molecules are hydrogen bonded. On adding acetone, its molecules get in between the host molecules and break some of the hydrogen bonds between them. Due to weakening of interactions, the solution shows positive deviation from Raoult's law [Fig. 2.6 (a)]. In a solution formed by adding carbon disulphide to acetone, the dipolar interactions between solute-solvent molecules are weaker than the respective interactions among the solute-solute and solvent-solvent molecules. This solution also shows positive deviation.

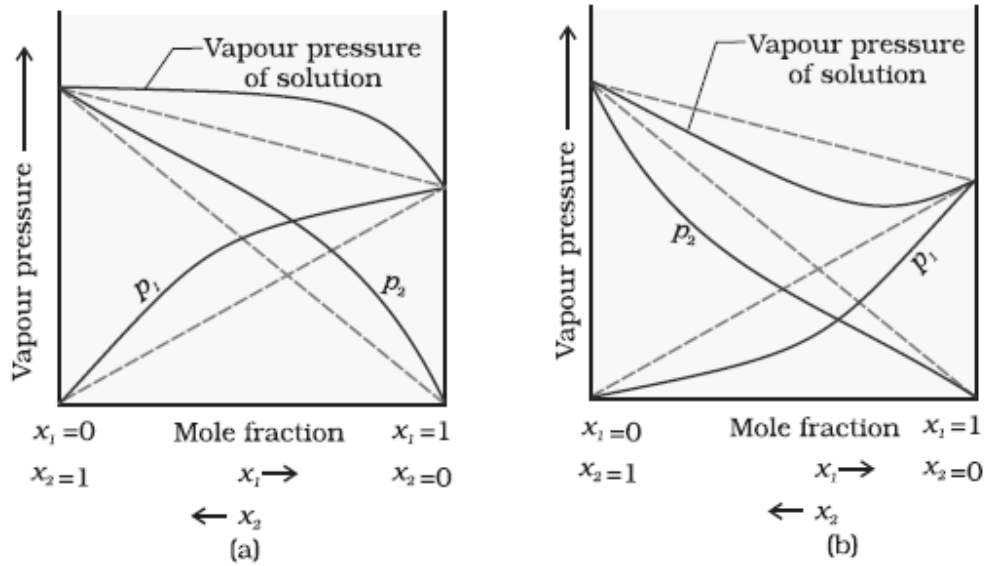
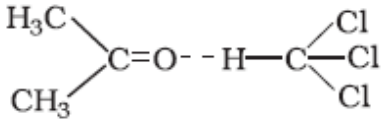


Fig.2.6 The vapour pressures of two component systems as a function of composition (a) a solution that shows positive deviation from Raoult's law and (b) a solution that shows negative deviation from Raoult's law.

In case of negative deviations from Raoult's law, the intermolecular attractive forces between A-A and B-B are weaker than those between A-B and leads to decrease in vapour pressure resulting in negative deviations. An example of this type is a mixture of phenol and aniline. In this case the intermolecular hydrogen bonding between phenolic proton and lone pair on nitrogen atom of aniline is stronger than the respective intermolecular hydrogen bonding between similar molecules. Similarly, a mixture of chloroform and acetone forms a solution with negative deviation from Raoult's law. This is because chloroform molecule is able to form hydrogen bond with acetone molecule as shown.



This decreases the escaping tendency of molecules for each component and consequently the vapour pressure decreases resulting in negative deviation from Raoult's law [Fig. 2.6. (b)].

Some liquids on mixing, form **azeotropes** which are binary mixtures having the same composition in liquid and vapour phase and boil at a constant temperature. In such cases, it is not possible to separate the components by fractional distillation. There are two types of azeotropes called **minimum boiling azeotrope and maximum boiling azeotrope**. The solutions which show a large positive deviation from Raoult's law form minimum boiling azeotrope at a specific composition. For example, ethanol-water mixture (obtained by fermentation of sugars) on fractional distillation gives a solution containing approximately 95% by volume of ethanol. Once this composition, known as azeotrope composition, has been achieved, the liquid and vapour have the same composition, and no further separation occurs.

The solutions that show large negative deviation from Raoult's law form maximum boiling azeotrope at a specific composition. Nitric acid and water is an example of this class of

azeotrope. This azeotrope has the approximate composition, 68% nitric acid and 32% water by mass, with a boiling point of 393.5 K.

2.6 Colligative Properties and Determination of Molar Mass

We have learnt in Section 2.4.3 that the vapour pressure of solution decreases when a non-volatile solute is added to a volatile solvent. There are many properties of solutions which are connected with this decrease of vapour pressure. These are: (1) relative lowering of vapour pressure of the solvent (2) depression of freezing point of the solvent (3) elevation of boiling point of the solvent and (4) osmotic pressure of the solution. **All these properties depend on the number of solute particles irrespective of their nature relative to the total number of particles present in the solution. Such properties are called *colligative properties*** (colligative: from Latin: co means together, ligare means to bind). In the following Sections we will discuss these properties one by one.

2.6.1 Relative Lowering of Vapour Pressure

We have learnt in Section 2.4.3 that the vapour pressure of a solvent in solution is less than that of the pure solvent. Raoult established that the lowering of vapour pressure depends only on the concentration of the solute particles and it is independent of their identity. The equation (2.20) given in Section 2.4.3 establishes a relation between vapour pressure of the solution, mole fraction and vapour pressure of the solvent, i.e.,

$$p_1 = x_1 p_1^0 \qquad (2.22)$$

The reduction in the vapour pressure of solvent (Δp_1) is given as:

$$\Delta p_1 = p_1^0 - p_1 = p_1^0 - p_1^0 x_1$$

$$= p_1^0 (1 - x_1) \qquad (2.23)$$

Knowing that $x_2 = 1 - x_1$, equation (2.23) reduces to

$$\Delta p_1 = x_2 p_1^0 \qquad (2.24)$$

In a solution containing several non-volatile solutes, the lowering of the vapour pressure depends on the sum of the mole fraction of different solutes.

Equation (2.24) can be written as

$$\frac{\Delta p_1}{p_1^0} = \frac{p_1^0 - p_1}{p_1^0} = x_2 \qquad (2.25)$$

The expression on the left hand side of the equation as mentioned earlier is called **relative lowering of vapour pressure and is equal to the mole fraction of the solute**. The above equation can be written as:

$$\frac{p_1^0 - p_1}{p_1^0} = \frac{n_2}{n_1 + n_2} \quad \left(\text{since } x_2 = \frac{n_2}{n_1 + n_2} \right) \qquad (2.26)$$

Here n_1 and n_2 are the number of moles of solvent and solute respectively present in the solution. For dilute solutions $n_2 << n_1$, hence neglecting n_2 in the denominator we have

$$\frac{p_1^0 - p_1}{p_1^0} = \frac{n_2}{n_1} \qquad (2.27)$$

$$\text{or} \quad \frac{p_1^0 - p_1}{p_1^0} = \frac{w_2 \times M_1}{M_2 \times w_1} \qquad (2.28)$$

Here w_1 and w_2 are the masses and M_1 and M_2 are the molar masses of the solvent and solute respectively.

From this equation (2.28), knowing all other quantities, the molar mass of solute (M_2) can be calculated.

2.6.2 Elevation of Boiling Point

We have learnt in Unit 5, Class XI, that the vapour pressure of a liquid increases with increase of temperature. It boils at the temperature at which its vapour pressure is equal to the atmospheric pressure. For example, water boils at 373.15 K (100° C) because at this temperature the vapour pressure of water is 1.013 bar (1 atmosphere). We have also learnt in the last section that vapour pressure of the solvent decreases in the presence of non-volatile solute. Fig. 2.7 depicts the variation of vapour pressure of the pure solvent and solution as a function of temperature. For example, the vapour pressure of an aqueous solution of sucrose is less than 1.013 bar at 373.15 K. In order to make this solution boil, its vapour pressure must be increased to 1.013 bar by raising the temperature above the boiling temperature of the pure solvent (water). Thus, the boiling point of a solution is always higher than that of the boiling point of the pure solvent in which the solution is prepared as shown in Fig. 2.7. Similar to lowering of vapour pressure, the elevation of boiling point also depends on the number of solute molecules rather than their nature. A solution of 1 mol of sucrose in 1000 g of water boils at 373.52 K at one atmospheric pressure.

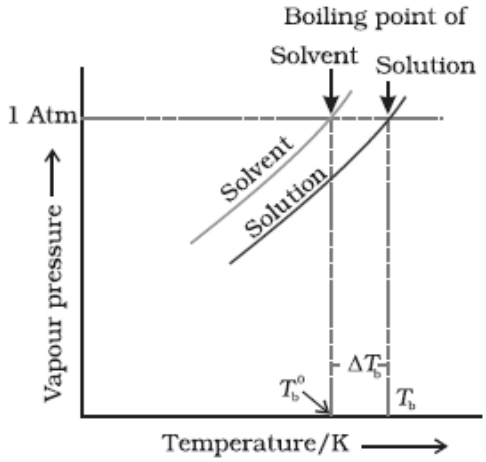


Fig. 2.7 The vapour pressure curve for solution lies below the curve for pure water. The diagram shows that ΔT_b denotes the elevation of boiling point of a solvent in solution.

Experiments have shown that for **dilute solutions** the elevation of boiling point (T_b) is directly proportional to the molal concentration of the solute in a solution. Thus

$$T_b \propto m \qquad (2.29)$$

$$\text{or} \quad T_b = K_b\, m \qquad (2.30)$$

Here m (molality) is the number of moles of solute dissolved in 1 kg of solvent and the constant of proportionality, K_b is called **Boiling Point Elevation Constant or Molal Elevation Constant (Ebullioscopic Constant)**. The unit of K_b is K kg mol^{-1}. Values of K_b for some common solvents are given in Table 2.3. If w_2 gram of solute of molar mass M_2 is dissolved in w_1 gram of solvent, then molality, m of the solution is given by the expression:

$$m = \frac{w_2/M_2}{w_1/1000} = \frac{1000 \times w_2}{M_2 \times w_1} \qquad (2.31)$$

Substituting the value of molality in equation (2.30) we get

$$\Delta T_b = \frac{k_b \times 1000 \times w_2}{M_2 \times w_1} \quad (2.32)$$

$$M_2 = \frac{1000 \times w_2 \times K_b}{\Delta T_b \times w_1} \quad (2.33)$$

Thus, in order to determine M_2, molar mass of the solute, known mass of solute in a known mass of the solvent is taken and ΔT_b is determined experimentally for a known solvent whose K_b value is known.

2.6.3 Depression of Freezing Point

The lowering of vapour pressure of a solution causes a lowering of the freezing point compared to that of the pure solvent (Fig. 2 8). We know that at the freezing point of a substance, the solid phase is in dynamic equilibrium with the liquid phase. Thus, the freezing point of a substance may be defined as the temperature at which the vapour pressure of the substance in its liquid phase is equal to its vapour pressure in the solid phase. A solution will freeze when its vapour pressure equals the vapour pressure of the pure solid solvent as is clear from Fig. 2.8. According to Raoult's law, when a non-volatile solid is added to the solvent its vapour pressure decreases and now it would become equal to that of solid solvent at lower temperature. Thus, the freezing point of the solvent decreases.

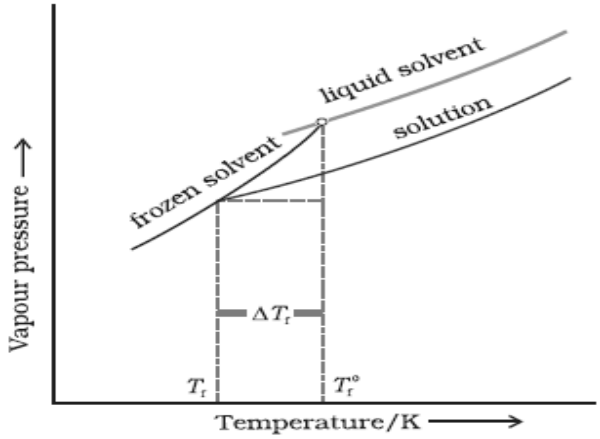


Fig. 2.8 Diagram showing ΔT_f, depression of the freezing point of a solvent in a solution.

Let T_f^0 be the freezing point of pure solvent and T_f be its freezing point when non-volatile solute is dissolved in it. The decrease in freezing point.

$\Delta T = T_f^0 - T_f$ is known as depression in freezing point.

Similar to elevation of boiling point, depression of freezing point (ΔT_f) for **dilute solution** (ideal solution) is directly proportional to molality, m of the solution. Thus,

$$\Delta T_f \propto m$$

$$\text{or} \quad \Delta T_f = K_f\, m \quad (2.34)$$

The proportionality constant, K_f, which depends on the nature of the solvent is known as **Freezing Point Depression Constant or Molal Depression Constant or Cryoscopic Constant**. The unit of K_f is K kg mol^{-1}. Values of K_f for some common solvents are listed in Table 2.3.

If w_2 gram of the solute having molar mass as M_2, present in w_1 gram of solvent, produces the depression in freezing point ΔT_f of the solvent then molality of the solute is given by the equation (2.31).

$$m = \frac{w_2 \ / \ M_2}{w_1/1000} \qquad (2.31)$$

Substituting this value of molality in equation (2.34) we get:

$$\Delta T_f = \frac{K_f \times w_2 / M_2}{w_1 / 1000}$$

$$\Delta T_f = \frac{K_f \times w_2 \times 1000}{M_2 \times w_1} \qquad (2.35)$$

$$M_2 = \frac{K_f \times w_2 \times 1000}{\Delta T_f \times w_1} \qquad (2.36)$$

Thus for determining the molar mass of the solute we should know the quantities w_1, w_2, ΔT_f, along with the molal freezing point depression constant.

The values of K_f and K_b, which depend upon the nature of the solvent, can be ascertained from the following relations.

$$K_f = \frac{R \times M_1 \times T_f^2}{1000 \times \Delta_{fus} H} \qquad (2.37)$$

$$K_b = \frac{R \times M_1 \times T_b^2}{1000 \times \Delta_{vap} H} \qquad (2.38)$$

Here the symbols R and $M1$ stand for the gas constant and molar mass of the solvent, respectively and T_f and T_b denote the freezing point and the boiling point of the pure solvent respectively in kelvin. Further, $\Delta_{fus}H$ and $\Delta_{vap}H$ represent the enthalpies for the fusion and vapourisation of the solvent, respectively.

2.6.4 Osmosis and Osmotic Pressure

There are many phenomena which we observe in nature or at home. For example, raw mangoes shrivel when pickled in brine (salt water); wilted flowers revive when placed in fresh water, blood cells collapse when suspended in saline water, etc. If we look into these processes we find one thing common in all, that is, all these substances are bound by membranes.

Table 2.3: Molal Boiling Point Elevation and Freezing Point Depression Constants for Some Solvents

Solvent	b. p./K	K_b /K kg mol^{-1}	f. p./K	K_f /K kg mol^{-1}
Water	373.15	0.52	273.0	1.86
Ethanol	351.5	1.20	155.7	1.99
Cyclohexane	353.74	2.79	279.55	20.00
Benzene	353.3	2.53	278.6	5.12
Chloroform	334.4	3.63	209.6	4.79
Carbon tetrachloride	350.0	5.03	250.5	31.8
Carbon disulphide	319.4	2.34	164.2	3.83
Diethyl ether	307.8	2.02	156.9	1.79
Acetic acid	391.1	2.93	290.0	3.90

These membranes can be of animal or vegetable origin and these occur naturally such as pig's bladder or parchment or can be synthetic such as cellophane. These membranes appear to be continuous sheets or films, yet they contain a network of submicroscopic holes or pores. Small solvent molecules, like water, can pass through these holes but the passage of bigger molecules like solute is hindered. Membranes having this kind of properties are known as *semipermeable membranes* (SPM).

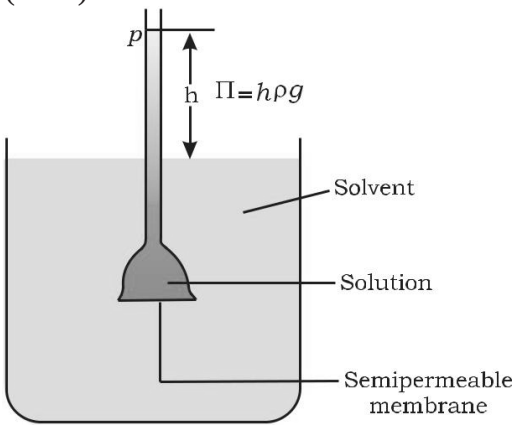


Fig. 2.9 Level of solution rises in the thistle funnel due to osmosis of solvent.

Assume that only solvent molecules can pass through these semi-permeable membranes. If this membrane is placed between the solvent and solution as shown in Fig. 2.9, the solvent molecules will flow through the membrane from pure solvent to the solution. **This process of flow of the solvent is called *osmosis*.**

The flow will continue till the equilibrium is attained. The flow of the solvent from its side to solution side across a semi-permeable membrane can be stopped if some extra pressure is applied on the solution. **This pressure that just stops the flow of solvent is called *osmotic pressure* of the solution.**

The flow of solvent from dilute solution to the concentrated solution across a semi-permeable membrane is due to osmosis. The important point to be kept in mind is that solvent molecules always flow from lower concentration to higher concentration of solution. The osmotic pressure has been found to depend on the concentration of the solution.

The osmotic pressure of a solution is the excess pressure that must be applied to a solution to prevent osmosis, i.e., to stop the passage of solvent molecules through a semi-permeable membrane into the solution. This is illustrated in Fig. 2.10. Osmotic pressure is a colligative property as it depends on the number of solute molecules and not on their identity. For dilute solutions, it has been found experimentally that **osmotic pressure is proportional to the molarity, *C* of the solution at a given temperature *T***. Thus:

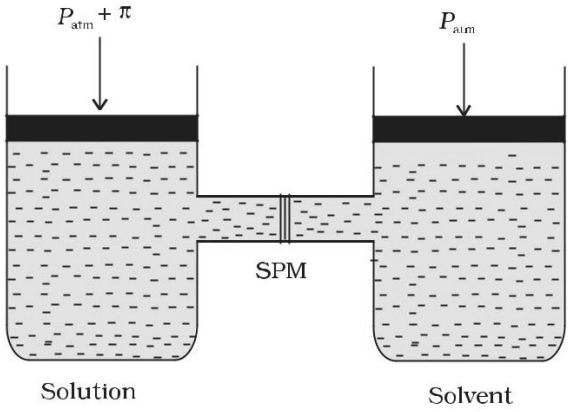


Fig. 2.10 The excess pressure equal to the osmotic pressure must be applied on the solution side to prevent osmosis.

$$\Pi = CRT \quad (2.39)$$

Here is Π the osmotic pressure and R is the gas constant.

$$\Pi = (n_2/V)RT \quad (2.40)$$

Here V is volume of a solution in litres containing n_2 moles of solute. If w_2 grams of solute, of molar mass, M_2 is present in the solution, then $n_2 = w_2 / M_2$ and we can write,

Thus, knowing the quantities w_2, T, and V we can calculate the molar mass of the solute.

Measurement of osmotic pressure provides another method of determining molar masses of solutes. This method is widely used to determine molar masses of proteins, polymers and other macromolecules. The osmotic pressure method has the advantage over other methods as pressure measurement is around the room temperature and the molarity of the solution is used instead of molality. As compared to other colligative properties, its magnitude is large even for very dilute solutions. The technique of osmotic pressure for determination of molar mass of solutes is particularly useful for biomolecules as they are generally not stable at higher temperatures and polymers have poor solubility.

Two solutions having same osmotic pressure at a given temperature are called isotonic solutions. When such solutions are separated by semi-permeable membrane no osmosis occurs between them. For example, the osmotic pressure associated with the fluid inside the blood cell is equivalent to that of 0.9% (mass/ volume) sodium chloride solution, called normal saline solution and it is safe to inject intravenously. On the other hand, if we place the cells in a solution containing more than 0.9% (mass/volume) sodium chloride, water will flow out of the cells and they would shrink. Such a solution is called **hypertonic**. If the salt concentration is less than 0.9% (mass/volume), the solution is said to be **hypotonic**. In this case, water will flow into the cells if placed in this solution and they would swell.

The phenomena mentioned in the beginning of this section can be explained on the basis of osmosis. A raw mango placed in concentrated salt solution loses water via osmosis and shrivel into pickle. Wilted flowers revive when placed in fresh water. A carrot that has become limp because of water loss into the atmosphere can be placed into the water making it firm once again. Water will move into them through osmosis. When placed in water containing less than 0.9% (mass/ volume) salt, blood cells collapse due to loss of water by osmosis. People taking a lot of

salt or salty food experience water retention in tissue cells and intercellular spaces because of osmosis. The resulting puffiness or swelling is called **edema**. Water movement from soil into plant roots and subsequently into upper portion of the plant is partly due to osmosis. The preservation of meat by salting and of fruits by adding sugar protects against bacterial action. Through the process of osmosis, a bacterium on salted meat or candid fruit loses water, shrivels and dies.

2.6.5 Reverse Osmosis and Water Purification

The direction of osmosis can be reversed if a pressure larger than the osmotic pressure is applied to the solution side. That is, now the pure solvent flows out of the solution through the semi permeable membrane. This phenomenon is called **reverse osmosis** and is of great practical utility. Reverse osmosis is used in desalination of sea water. A schematic set up for the process is shown in Fig. 2.11.

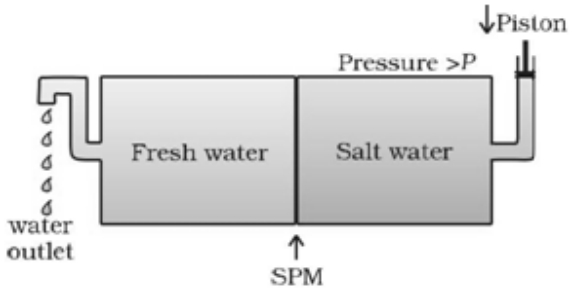


Fig. 2.11 **Reverse osmosis occurs when a pressure larger than the osmotic pressure is applied to the solution.**

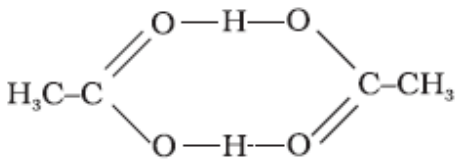
When pressure more than osmotic pressure is applied, pure water is squeezed out of the sea water through the membrane. A variety of polymer membranes are available for this purpose.

The pressure required for the reverse osmosis is quite high. A workable porous membrane is a film of cellulose acetate placed over a suitable support. Cellulose acetate is permeable to water but impermeable to impurities and ions present in sea water. These days many countries use desalination plants to meet their potable water requirements.

2. 7 Abnormal Molar Masses

We know that ionic compounds when dissolved in water dissociate into cations and anions. For example, if we dissolve one mole of KCl (74.5 g) in water, we expect one mole each of K^+ and Cl^- ions to be released in the solution. If this happens, there would be two moles of particles in the solution. If we ignore interionic attractions, one mole of KCl in one kg of water would be expected to increase the boiling point by 2×0.52 K = 1.04 K. Now if we did not know about the degree of dissociation, we could be led to conclude that the mass of 2 mol particles is 74.5 g and the mass of one mole of KCl would be 37.25 g. This brings into light the rule that, when there is dissociation of solute into ions, the experimentally determined molar mass is always lower than the true value.

$$2\ CH_3COOH \rightleftharpoons (CH_3COOH)_2$$



Molecules of ethanoic acid (acetic acid) dimerise in benzene due to hydrogen bonding. This normally happens in solvents of low dielectric constant. In this case the number of particles is reduced due to dimerisation. Association of molecules is depicted as follows:

It can be undoubtedly stated here that if all the molecules of ethanoic acid associate in benzene, then ΔT_b or ΔT_f for ethanoic acid will be half of the normal value. The molar mass calculated on the basis of this ΔT_b or ΔT_f will, therefore, be twice the expected value. Such a molar mass that is either lower or higher than the expected or normal value is called as **abnormal molar mass.**

In 1880 van't Hoff introduced a factor *i*, known as the van't Hoff factor, to account for the extent of dissociation or association. This factor *i* is defined as:

$$i = \frac{\text{Normal molar mass}}{\text{Abnormal molar mass}}$$

$$= \frac{\text{Observed colligative property}}{\text{Calculated colligative property}}$$

$$i = \frac{\text{Total number of moles of particles after association/dissociation}}{\text{Number of moles of particles before association/dissociation}}$$

Here abnormal molar mass is the experimentally determined molar mass and calculated colligative properties are obtained by assuming that the non-volatile solute is neither associated nor dissociated. In case of association, value of *i* is less than unity while for dissociation it is greater than unity. For example, the value of *i* for aqueous KCl solution is close to 2, while the value for ethanoic acid in benzene is nearly 0.5.

Inclusion of van't Hoff factor modifies the equations for colligative properties as follows:

Relative lowering of vapour pressure of solvent,

$$\frac{p_1^o - p_1}{p_1^o} = i.\frac{n_2}{n_1}$$

Elevation of Boiling point, T_b $= i\,K_b\,m$

Depression of Freezing point, T_f $= i\,K_f\,m$

Osmotic pressure of solution, ð $= i\,n_2\,R\,T\,/\,V$

Table 2.4 depicts values of the factor, *i* for several strong electrolytes. For KCl, NaCl and $MgSO_4$, *i* approach 2 as the solution becomes very dilute. As expected, the value of *i* gets close to 3 for K_2SO_4.

Table 2.4: Values of Van't Hoff factor, *i*, at Various Concentrations for NaCl, KCl, $MgSO_4$ and K_2SO_4.

Salt	Values of *i** 0.1 m	0.01 m	0.001 m	van't Hoff Factor *i* for complete dissociation of solute
NaCl	1.87	1.94	1.97	2.00
KCl	1.85	1.94	1.98	2.00
$MgSO_4$	1.21	1.53	1.82	2.00
K_2SO_4	2.32	2.70	2.84	3.00

* *represent* i *values for incomplete dissociation.*

Summary

A solution is a homogeneous mixture of two or more substances. Solutions are classified as solid, liquid and gaseous solutions. The concentration of a solution is expressed in terms of mole fraction, molarity, molality and in percentages. The dissolution of a gas in a liquid is governed by **Henry's law**, according to which, at a given temperature, the **solubility of a gas in a liquid is directly proportional to the partial pressure of the gas**. The vapour pressure of the solvent is lowered by the presence of a non-volatile solute in the solution and this lowering of vapour pressure of the solvent is governed by Raoult's law, according to which the **relative lowering of vapour pressure of the solvent over a solution is equal to the mole fraction of a non-volatile solute present in the solution**. However, in a binary liquid solution, if both the components of the solution are volatile then another form of Raoult's law is used. Mathematically, this form of the Raoult's law is stated as: $p_{total} = p_1^0 x_1 + p_2^0 x_2$. **Solutions which obey Raoult's law over the entire range of concentration are called ideal solutions**. Two types of deviations from Raoult's law, called positive and negative deviations are observed. Azeotropes arise due to very large deviations from Raoult's law.

The properties of solutions which depend on the number of solute particles and are independent of their chemical identity are called colligative properties. These are lowering of vapour pressure, elevation of boiling point, depression of freezing point and osmotic pressure. The process of osmosis can be reversed if a pressure higher than the osmotic pressure is applied to the solution. Colligative properties have been used to determine the molar mass of solutes. Solutes which dissociate in solution exhibit molar mass lower than the actual molar mass and those which associate show higher molar mass than their actual values.

Quantitatively, the extent to which a solute is dissociated or associated can be expressed by van't Hoff factor *i*. This factor has been defined as ratio of normal molar mass to experimentally determined molar mass or as the ratio of observed colligative property to the calculated colligative property.

IMPORTANT QUESTIONS

1) State Raoult's law and Henry's law.
2) Define Osmatic pressure
3) What are isotonic solutions
4) Calculate the molefraction of ethylene glycol in a solution containing 20% of ethylene glycol by mass.
5) Vapour pressure of water at 293Kis 17.535mmHg.Calculate the vapour pressure of the solution at 293Kwhen 25g of glucose is dissolved in 450g of water.

Chapter 3
ELECTROCHEMISTRY AND CHEMICAL KINETICS
ELECTROCHEMISTRY

Electrochemistry is the study of production of electricity from energy released during spontaneous chemical reactions and the use of electrical energy to bring about non-spontaneous chemical transformations. The subject is of importance both for theoretical and practical considerations. A large number of metals, sodium hydroxide, chlorine, fluorine and many other chemicals are produced by electrochemical methods. Batteries and fuel cells convert chemical energy into electrical energy and are used on a large scale in various instruments and devices. The reactions carried out electrochemically can be energy efficient and less polluting. Therefore, study of electrochemistry is important for creating new technologies that are eco-friendly. The transmission of sensory signals through cells to brain and vice versa and communication between the cells are known to have electrochemical origin. Electrochemistry, is therefore, a very vast and interdisciplinary subject. In this Unit, we will cover only some of its important elementary aspects.

3 . 1 Electrochemical Cells

The construction and functioning of **Daniell cell** (Fig. 3.1). This cell converts the chemical energy liberated during the redox reaction

$Zn(s) + Cu^{2+}(aq) \rightarrow Zn^{2+}(aq) + Cu(s)$

to electrical energy and has an electrical potential equal to 1.1 V when concentration of Zn^{2+} and Cu^{2+} ions is unity (1 mol dm^{-3})*. Such a device is called a **galvanic** or a **voltaic** cell (Fig.3.2).

If an external opposite potential is applied and increased slowly, we find that the reaction continues to take place till the opposing voltage reaches the value 1.1 V when, the reaction stops altogether and no current flows through the cell. Any further increase in the external potential again starts the reaction but in the opposite direction. It now functions as an **electrolytic cell**, a device for using electrical energy to carry non-spontaneous chemical reactions. Both types of cells are quite important and we shall study some of their salient features in the following pages.

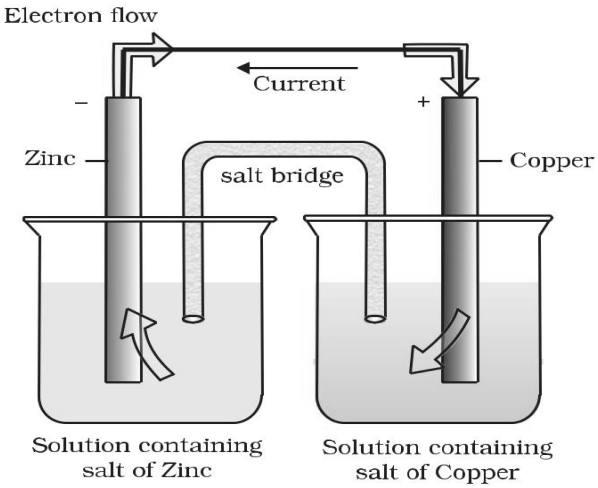


Fig.3.1 Daniell cell having electrodes of zinc and copper dipping in the solutions of their respective salts.

---

* Strictly speaking activity should be used instead of concentration. It is directly proportional to concentration. In dilute solutions, it is equal to concentration. You will study more about it in higher classes.

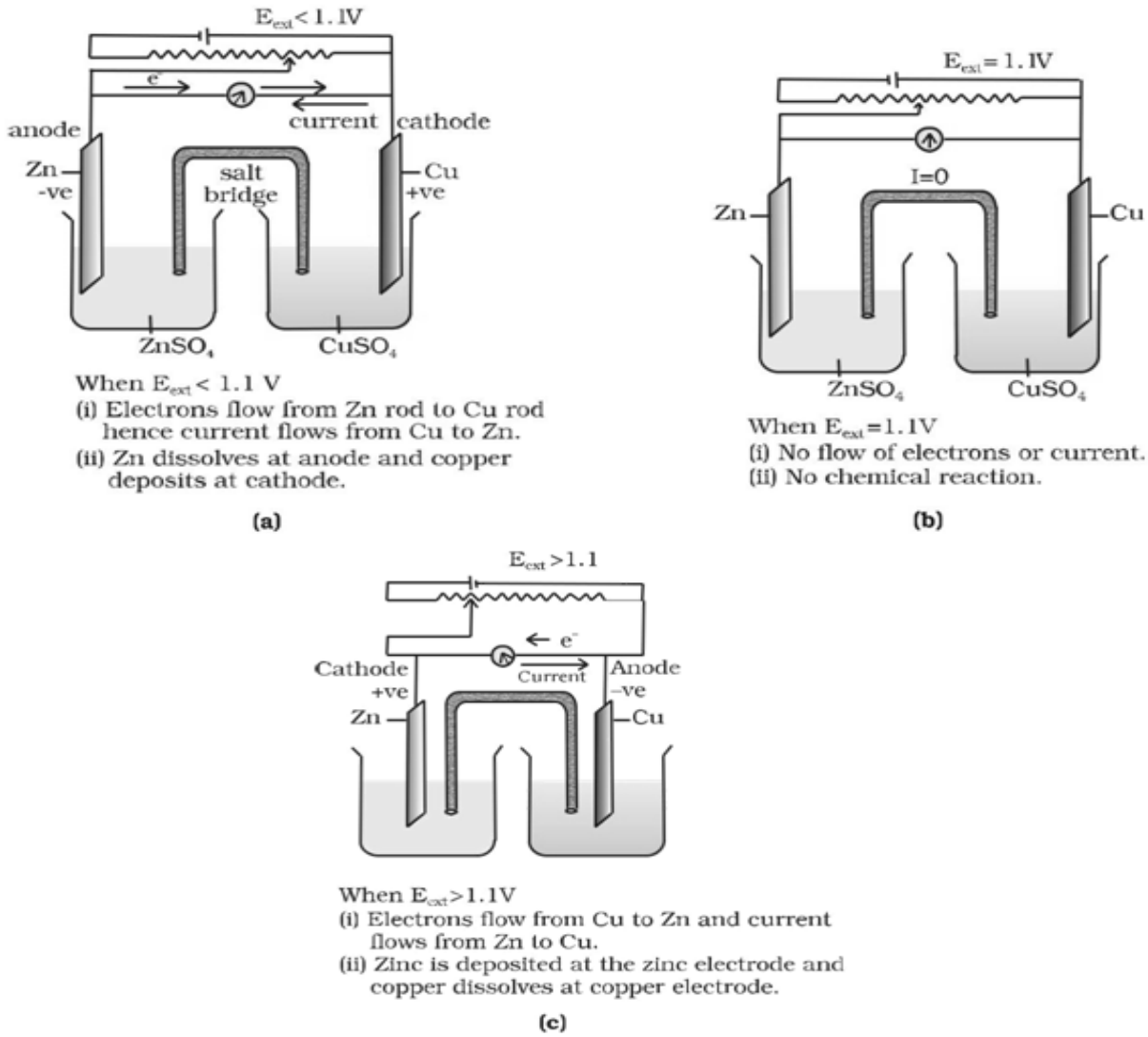


Fig. 3.2 Functioning of Daniell cell when external voltage E_{ext} opposing the cell potential is applied.

3.2 Galvanic Cells

As mentioned earlier a galvanic cell is an electrochemical cell that converts the chemical energy of a spontaneous redox reaction into electrical energy. In this device the Gibbs energy of the spontaneous redox reaction is converted into electrical work which may be used for running motor or other electrical gadgets like heater, fan, geyser, etc.

Daniell cell discussed earlier is one such cell in which the following redox reaction occurs.

$$Zn(s) + Cu^{2+}(aq) \rightarrow Zn^{2+}(aq) + Cu(s)$$

This reaction is a combination of two half reactions whose addition gives the overall cell reaction:

(i) $Cu^{2+} + 2e^- \rightarrow Cu(s)$ (reduction half reaction)

(ii) $Zn(s) \rightarrow Zn^{2+} + 2e^-$ (oxidation half reaction)

These reactions occur in two different portions of the Daniell cell. The reduction half reaction occurs on the copper electrode while the oxidation half reaction occurs on the zinc electrode. These two portions of the cell are also called **half-cells** or **redox couples**. The copper electrode may be called the reduction half cell and the zinc electrode, the oxidation half-cell.

We can construct innumerable number of galvanic cells on the pattern of Daniell cell by taking combinations of different half-cells. Each half-cell consists of a metallic electrode dipped into an electrolyte. The two half-cells are connected by a metallic wire through a voltmeter and a switch externally. The electrolytes of the two half-cells are connected internally through a salt bridge as shown in Fig. 3.1. Sometimes, both the electrodes dip in the same electrolyte solution and in such cases we don't require a salt bridge.

At each electrode-electrolyte interface there is a tendency of metal ions from the solution to deposit on the metal electrode trying to make it positively charged. At the same time, metal atoms of the electrode have a tendency to go into the solution as ions and leave behind the electrons at the electrode trying to make it negatively charged. At equilibrium, there is a separation of charges and depending on the tendencies of the two opposing reactions, the electrode may be positively or negatively charged with respect to the solution. A potential difference develops between the electrode and the electrolyte which is called electrode potential. When the concentrations of all the species involved in a half-cell is unity then the electrode potential is known as standard electrode potential. According to IUPAC convention, standard reduction potentials are now called standard electrode potentials. In a galvanic cell, the half-cell in which oxidation takes place is called anode and it has a negative potential with respect to the solution. The other half-cell in which reduction takes place is called cathode and it has a positive potential with respect to the solution. Thus, there exists a potential difference between the two electrodes and as soon as the switch is in the *on* position the electrons flow from negative electrode to positive electrode. The direction of current flow is opposite to that of electron flow.

The potential difference between the two electrodes of a galvanic cell is called the *cell potential* and is measured in volts. The *cell potential* is the difference between the electrode potentials (reduction potentials) of the cathode and anode. It is called the *cell electromotive force (emf)* of the cell when no current is drawn through the cell. It is now an accepted convention that we keep the anode on the left and the cathode on the right while representing the galvanic cell. A galvanic cell is generally represented by putting a vertical line between metal and electrolyte solution and putting a double vertical line between the two electrolytes connected by a salt bridge. Under this convention the emf of the cell is positive and is given by the potential of the half-cell on the right hand side minus the potential of the half-cell on the left hand side i.e.

$$E_{cell} = E_{right} - E_{left}$$

This is illustrated by the following example:

Cell reaction:

$$Cu(s) + 2Ag^{+}(aq) \rightarrow Cu^{2+}(aq) + 2\,Ag(s) \qquad (3.4)$$

Half-cell reactions:

Cathode (*reduction*): $2Ag^{+}(aq) + 2e^{-} \rightarrow 2Ag(s)$ (3.5)

Anode (*oxidation*): $Cu(s) \rightarrow Cu^{2+}(aq) + 2e^{-}$ (3.6)

It can be seen that the sum of Anode (oxidation) and Cathode (reduction) leads to over all Cell reaction in the cell and that silver electrode acts as a cathode and copper electrode acts as an anode. The cell can be represented as:

$$Cu(s)|Cu^{2+}(aq)||Ag^{+}(aq)|Ag(s)$$

and we have $E_{cell} = E_{right} - E_{left} = E_{Ag^+\,Ag} - E_{Cu^{2+}\,Cu}$

Measurement of Electrode Potential

The potential of individual half-cell cannot be measured. We can measure only the difference between the two half-cell potentials that gives the emf of the cell. If we arbitrarily choose the potential of one electrode (half- cell) then that of the other can be determined with respect to this. According to convention, a half-cell called standard hydrogen electrode represented by Pt(s) $H_2(g)$ $H^+(aq)$, is assigned a zero potential at all temperatures corresponding to the reaction

$$H^+(aq) + e^- \rightarrow \frac{1}{2} H_2(g)$$

The standard hydrogen electrode consists of a platinum electrode coated with platinum black (Fig.3.3). The electrode is dipped in an acidic solution and pure hydrogen gas is bubbled through it. The concentration of both the reduced and oxidised forms of hydrogen is maintained at unity. This implies that the pressure of hydrogen gas is one bar and the concentration of hydrogen ion in the solution is one molar.

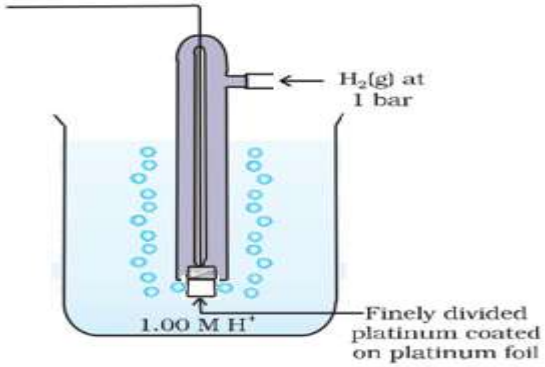


Fig. 3.3 Standard Hydrogen Electrode (SHE).

At 298 K the emf of the cell, standard hydrogen electrode second half-cell constructed by taking standard hydrogen electrode as anode (reference half-cell) and the other half-cell as cathode, gives the reduction potential of the other half-cell. If the concentrations of the oxidised and the reduced forms of the species in the right hand half-cell are unity, then the cell potential is equal to standard electrode potential, E_R of the given half-cell.

$E = E_R - E_L$

As E_L for standard hydrogen electrod is zero.

$E = E_R - 0 = E_R$

The measured emf of the cell :

Pt(s) H_2(g, 1 bar) H^+ (aq, 1 M) Cu^{2+} (aq, 1 M) Cu

is 0.34 V and it is also the value for the standard electrode potential of the half-cell corresponding to the reaction :

Cu^{2+} (aq, 1M) + 2 e^- → Cu(s)

Similarly, the measured emf of the cell :

Pt(s) H_2(g, 1 bar) H^+ (aq, 1 M) Zn^{2+} (aq, 1M) Zn

is -0.76 V corresponding to the standard electrode potential of the half-cell reaction:

Zn^{2+} (aq, 1 M) + $2e^-$ → Zn(s)

The positive value of the standard electrode potential in the first case indicates that Cu^{2+} ions get reduced more easily than H^+ ions. The reverse process cannot occur, that is, hydrogen ions cannot oxidise Cu (or alternatively we can say that hydrogen gas can reduce copper ion) under the standard conditions described above. Thus, Cu does not dissolve in HCl. In nitric acid it is oxidised by nitrate ion and not by hydrogen ion. The negative value of the standard electrode potential in the second case indicates that hydrogen ions can oxidise zinc (or zinc can reduce hydrogen ions).

In view of this convention, the half reaction for the Daniell cell in
The above figure can be written as:

$$\text{Left electrode : } Zn(s) \rightarrow Zn^{2+}(aq, 1\ M) + 2\ e^-$$
$$\text{Right electrode: } Cu^{2+}(aq, 1\ M) + 2\ e^- \rightarrow Cu(s)$$

The overall reaction of the cell is the sum of above two reactions and we obtain the equation:

$$Zn(s) + Cu^{2+}(aq) \rightarrow Zn^{2+}(aq) + Cu(s)$$
$$\text{Emf of the cell} = E_{cell}^0 = E^0_R - E^0_L$$
$$= 0.34V - (-0.76)V = 1.10\ V$$

Sometimes metals like platinum or gold are used as inert electrodes. They do not participate in the reaction but provide their surface for oxidation or reduction reactions and for the conduction of electrons. For example, Pt is used in the following half-cells: Hydrogen electrode: $Pt(s)|H_2(g)|\ H^+(aq)$

$$\text{With half-cell reaction: } H^+(aq) + e^- \rightarrow \tfrac{1}{2}\ H_2(g)$$
$$\text{Bromine electrode: } Pt(s)|Br_2(aq)|\ Br^-(aq)$$
$$\text{With half-cell reaction: } \tfrac{1}{2}\ Br_2(aq) + e^- \rightarrow Br^-(aq)$$

The standard electrode potentials are very important and we can extract a lot of useful information from them. The values of standard electrode potentials for some selected half-cell reduction reactions are given in Table 3.1. If the standard electrode potential of an electrode is greater than zero, then its reduced form is more stable compared to hydrogen gas. Similarly, if the standard electrode potential is negative then hydrogen gas is more stable than the reduced form of the species. It can be seen that the standard electrode potential for fluorine is the highest in the Table indicating that fluorine gas (F_2) has the maximum tendency to get reduced to fluoride ions (F^-) and therefore fluorine gas is the strongest oxidising agent and fluoride ion is the weakest reducing agent. Lithium has the lowest electrode potential indicating that lithium ion is the weakest oxidising agent while lithium metal is the most powerful reducing agent in an aqueous solution. It may be seen that as we go from top to bottom in Table 3.1 the standard electrode potential decreases and with this, decreases the oxidising power of the species on the left and increases the reducing power of the species on the right hand side of the reaction. Electrochemical cells are extensively used for determining the pH of solutions, solubility product, equilibrium constant and other thermodynamic properties and for potentiometric titrations.

Table 3.1 The standard electrode potentials at 298 K

Ions are present as aqueous species and H_2O as liquid; gases and solids are shown by g and s.

Reaction (Oxidised form + ne^-	**→ Reduced form)**		***E* /V**
$F_2(g) + 2e^-$	$\rightarrow 2F^-$		2.87
$Co^{3+} + e^-$	$\rightarrow Co^{2+}$		1.81
$H_2O_2 + 2H^+ + 2e^-$	$\rightarrow 2H_2O$		1.78
$MnO_4^- + 8H^+ + 5e^-$	$\rightarrow Mn^{2+} + 4H_2O$		1.51
$Au^{3+} + 3e^-$	$\rightarrow$ Au(s)		1.40
$Cl_2(g) + 2e^-$	$\rightarrow 2Cl^-$		1.36
$Cr_2O_7^{2-} + 14H^+ + 6e^-$	$\rightarrow 2Cr^{3+} + 7H_2O$		1.33
$O_2(g) + 4H^+ + 4e^-$	$\rightarrow 2H_2O$		1.23
$MnO_2(s) + 4H^+ + 2e^-$	$\rightarrow Mn^{2+} + 2H_2O$		1.23
$Br_2 + 2e^-$	$\rightarrow 2Br^-$		1.09
$NO_3^- + 4H^+ + 3e^-$	$\rightarrow NO(g) + 2H_2O$		0.97
$2Hg^{2+} + 2e^-$	$\rightarrow Hg_2^{2+}$		0.92
$Ag^+ + e^-$	$\rightarrow$ Ag(s)		0.80
$Fe^{3+} + e^-$	$\rightarrow Fe^{2+}$		0.77
$O_2(g) + 2H^+ + 2e^-$	$\rightarrow H_2O_2$		0.68
$I_2 + 2e^-$	$\rightarrow 2I^-$		0.54
$Cu^+ + e^-$	$\rightarrow$ Cu(s)		0.52
$Cu^{2+} + 2e^-$	$\rightarrow$ Cu(s)		0.34
AgCl(s) + e^-	$\rightarrow$ Ag(s) + Cl^-		0.22
AgBr(s) + e^-	$\rightarrow$ Ag(s) + Br^-		0.10
$2H^+ + 2e^-$	**$\rightarrow H_2(g)$**		0.00
$Pb^{2+} + 2e^-$	$\rightarrow$ Pb(s)		-0.13
$Sn^{2+} + 2e^-$	$\rightarrow$ Sn(s)		-0.14
$Ni^{2+} + 2e^-$	$\rightarrow$ Ni(s)		-0.25
$Fe^{2+} + 2e^-$	$\rightarrow$ Fe(s)		-0.44
$Cr^{3+} + 3e^-$	$\rightarrow$ Cr(s)		-0.74
$Zn^{2+} + 2e^-$	$\rightarrow$ Zn(s)		-0.76
$2H_2O + 2e^-$	$\rightarrow H_2(g) + 2OH^-$ (aq)		-0.83
$Al^{3+} + 3e^-$	$\rightarrow$ Al(s)		-1.66
$Mg^{2+} + 2e^-$	$\rightarrow$ Mg(s)		-2.36
$Na^+ + e^-$	$\rightarrow$ Na(s)		-2.71
$Ca^{2+} + 2e^-$	$\rightarrow$ Ca(s)		-2.87
$K^+ + e^-$	$\rightarrow$ K(s)		-2.93
$Li^+ + e^-$	$\rightarrow$ Li(s)		-3.05

Increasing strength of reducing agent

1. A negative *E* means that the redox couple is a stronger reducing agent than the H^+/H_2 couple.
2. A positive *E* means that the redox couple is a weaker reducing agent than the H^+/H_2 couple.

3 . 3 Nernst Equation

We have assumed in the previous section that the concentration of all the species involved in the electrode reaction is unity. This need not be always true. Nernst showed that for the electrode reaction:

$$M^{n+}(aq) + ne^- \rightarrow M(s)$$

the electrode potential at any concentration measured with respect to standard hydrogen electrode can be represented by:

$$E_{(M^{n+}/M)} = E^{\ominus}_{(M^{n+}/M)} - \frac{RT}{nF} \ln \frac{[M]}{[M^{n+}]}$$

but concentration of solid M is taken as unity and we have

$$E_{(M^{n+}/M)} = E^{\ominus}_{(M^{n+}/M)} - \frac{RT}{nF} \ln \frac{1}{[M^{n+}]}$$

$E_{(M^{n+}/M)}$ has already been defined, R is gas constant (8.314 JK^{-1} mol^{-1}), F is Faraday constant (96487 C mol^{-1}), T is temperature in kelvin and $[M^{n+}]$ is the concentration of the species, M^{n+}.

In Daniell cell, the electrode potential for any given concentration of Cu^{2+} and Zn^{2+} ions, we write For Cathode:

$$E_{(Cu^{2+}/Cu)} = E^{\ominus}_{(Cu^{2+}/Cu)} - \frac{RT}{2F} \ln \frac{1}{[Cu^{2+}(aq)]}$$

For Anode:

$$E_{(Zn^{2+}/Zn)} = E^{\ominus}_{(Zn^{2+}/Zn)} - \frac{RT}{2F} \ln \frac{1}{[Zn^{2+}(aq)]}$$

The cell potential, $E_{(cell)} = E_{(Cu^{2+}/Cu)} - E_{(Zn^{2+}/Zn)}$

$$= E^{\ominus}_{(Cu^{2+}/Cu)} - \frac{RT}{2F} \ln \frac{1}{[Cu^{2+}(aq)]} - E^{\ominus}_{(Zn^{2+}/Zn)} + \frac{RT}{2F} \ln \frac{1}{[Zn^{2+}(aq)]}$$

$$= E^{\ominus}_{(Cu^{2+}/Cu)} - E^{\ominus}_{(Zn^{2+}/Zn)} - \frac{RT}{2F} \ln \frac{1}{[Cu^{2+}(aq)]} - \ln \frac{1}{[Zn^{2+}(aq)]}$$

$$E_{(cell)} = E^{\ominus}_{(cell)} - \frac{RT}{2F} \ln \frac{[Zn^{2+}]}{[Cu^{2+}]}$$

It can be seen that $E_{(cell)}$ depends on the concentration of both Cu^{2+} and Zn^{2+} ions. It increases with increase in the concentration of Cu^{2+} ions and decrease in the concentration of Zn^{2+} ions.

By converting the natural logarithm in the above Eq to the base 10 and substituting the values of R, F and T = 298 K, it reduces to

$$E_{(cell)} = E^{\ominus}_{(cell)} - \frac{0.059}{2} \log \frac{[Zn^{2+}]}{[Cu^{2+}]}$$

We should use the same number of electrons (n) for both the electrodes and thus for the following cell

$Ni(s) \mid Ni^{2+}(aq) \parallel Ag^{+}(aq) \mid Ag$

The cell reaction is $Ni(s) + 2Ag^{+}(aq) \rightarrow Ni^{2+}(aq) + 2Ag(s)$

The Nernst equation can be written as

$$E_{(cell)} = E^{\ominus}_{(cell)} - \frac{RT}{2F} \ln \frac{[Ni^{2+}]}{[Ag^{+}]^2}$$

and for a general electrochemical reaction of the type:

$$a\,A + bB \xrightarrow{ne} cC + dD$$

Nernst equation can be written as:

$$E_{(cell)} = E^{\ominus}_{(cell)} - \frac{RT}{nF} \ln Q$$

$$= E^{\ominus}_{(cell)} - \frac{RT}{nF} \ln \frac{[C]^c [D]^d}{[A]^a [B]^b}$$

Equilibrium Constant from Nernst Equation

If the circuit in Daniell cell (Fig. 3.1) is closed then we note that the reaction

$Zn(s) + Cu^{2+}(aq) \rightarrow Zn^{2+}(aq) + Cu(s)$

takes place and as time passes, the concentration of Zn^{2+} keeps on increasing while the concentration of Cu^{2+} keeps on decreasing. At the same time voltage of the cell as read on the voltmeter keeps on decreasing. After some time, we shall note that there is no change in the concentration of Cu^{2+} and Zn^{2+} ions and at the same time, voltmeter gives zero reading. This indicates that equilibrium has been attained. In this situation the Nernst equation may be written as:

$$E_{(cell)} = 0 = E^{\ominus}_{(cell)} - \frac{2.303RT}{2F} \log \frac{[Zn^{2+}]}{[Cu^{2+}]}$$

$$\text{or } E^{\ominus}_{(cell)} = \frac{2.303RT}{2F} \log \frac{[Zn^{2+}]}{[Cu^{2+}]}$$

But at equilibrium,

$\frac{[Zn^{2+}]}{[Cu^{2+}]} = K_c$ for the reaction 3.1

and at T = 298K the above equation can be written as

$$E^{\ominus}_{(cell)} = \frac{0.059\ V}{2} \log K_C = 1.1\ V \qquad (E^{\ominus}_{(cell)} = 1.1V)$$

$$\log K_C = \frac{(1.1V \times 2)}{0.059\ V} = 37.288$$

$K_C = 2 \times 10^{37}$ at 298K.

In general,

$$E^{\ominus}_{(cell)} = \frac{2.303RT}{nF} \log K_C$$

Thus, the above Equation gives a relationship between equilibrium constant of the reaction and standard potential of the cell in which that reaction takes place. Thus, equilibrium constants of the reaction, difficult to measure otherwise, can be calculated from the corresponding *E* value of the cell.

Electro- chemical Cell and Gibbs Energy of the Reaction

Electrical work done in one second is equal to electrical potential multiplied by total charge passed. If we want to obtain maximum work from a galvanic cell then charge has to be passed reversibly. The reversible work done by a galvanic cell is equal to decrease in its Gibbs energy and therefore, if the emf of the cell is *E* and *nF* is the amount of charge passed and $_rG$ is the Gibbs energy of the reaction, then

$$\Delta_r G = - nFE_{(cell)}$$

It may be remembered that *E*(cell) is an intensive parameter but Δr*G* is an extensive thermodynamic property and the value depends on *n*. Thus, if we write the reaction

$$Zn(s) + Cu^{2+}(aq) \longrightarrow Zn^{2+}(aq) + Cu(s)$$

$$\Delta_r G = -2FE_{(cell)}$$

but when we write the reaction

$$2\ Zn(s) + 2\,Cu^{2+}(aq) \longrightarrow 2\ Zn^{2+}(aq) + 2\,Cu(s)$$

$$\Delta_r G = -4FE_{(cell)}$$

If the concentration of all the reacting species is unity, then $E_{(cell)} = {}_{(cell)}E$V and we have

$$\Delta_r G^{\ominus} = -nF\,E^{\ominus}_{(cell)} \qquad (3.16)$$

Thus, from the measurement of $_{(cell)}E$V we can obtain an important thermodynamic quantity, Δr*G*_, standard Gibbs energy of the reaction. From the latter we can calculate equilibrium constant by the equation:

Δr*G*_ = –*RT* ln *K*.

3.4 Conductance of Electrolytic Solutions

It is necessary to define a few terms before we consider the subject of conductance of electricity through electrolytic solutions. The electrical resistance is represented by the symbol *'R'* and it is measured in ohm (Ω) which in terms of SI base units is equal to (kg m_2)/(s^3 A^2). It can be measured with the help of a Wheatstone bridge with which you are familiar from your study of physics. The electrical resistance of any object is directly proportional to its length, *l*, and inversely proportional to its area of cross section, *A*. That is,

$$R \propto \frac{l}{A} \text{ or } R = \rho\,\frac{l}{A}$$

The constant of proportionality, ρ (Greek, rho), is called resistivity (specific resistance). Its SI units are ohm metre (Ω m) and quite often its sub-multiple, ohm centimetre (Ω cm) is also used. IUPAC recommends the use of the term resistivity over specific resistance and hence in the rest of the book we shall use the term resistivity. Physically, the resistivity for a substance is its resistance when it is one metre long and its area of cross section is one m2. It can be seen that:

$$1\ \Omega\ m = 100\ \Omega\ cm \text{ or } 1\ \Omega\ cm = 0.01\ \Omega\ m$$

The inverse of resistance, *R*, is called **conductance**, *G*, and we have the relation:

$$G = \frac{1}{R} = \frac{A}{\rho l} = \kappa\frac{A}{l}$$

The SI unit of conductance is siemens, represented by the symbol 'S' and is equal to ohm^{-1} (also known as mho) or Ω^{-1}. The inverse of resistivity, called **conductivity** (specific conductance) is represented by the symbol, κ (Greek, kappa). IUPAC has recommended the use of term conductivity over specific conductance and hence we shall use the term conductivity in the rest of the book. The SI units of conductivity are S m^{-1} but quite often, κ is expressed in S cm^{-1}. Conductivity of a material in S m–1 is its conductance when it is 1 m long and its area of cross section is 1 m^2. It may be noted that 1 S cm^{-1} = 100 S m^{-1}.

Material	Conductivity/ S m^{-1}	Material	Conductivity/ S m^{-1}
Conductors		***Aqueous Solutions***	
Sodium	2.1×10^3	Pure water	3.5×10^{-5}
Copper	5.9×10^3	0.1 M HCl	3.91
Silver	6.2×10^3	0.01M KCl	0.14
Gold	4.5×10^3	0.01M NaCl	0.12
Iron	1.0×10^3	0.1 M HAc	0.047
Graphite	1.2×10	0.01M HAc	0.016
Insulators		***Semiconductors***	
Glass	1.0×10^{-16}	CuO	1×10^{-7}
Teflon	1.0×10^{-18}	Si	1.5×10^{-2}
		Ge	2.0

It can be seen from Table 3.2 that the magnitude of conductivity varies a great deal and depends on the nature of the material. It also depends on the temperature and pressure at which the measurements are made. Materials are classified into conductors, insulators and semiconductors depending on the magnitude of their conductivity. Metals and their alloys have very large conductivity and are known as conductors. Certain non-metals like carbon-black, graphite and some organic polymers* are also electronically conducting. Substances like glass, ceramics, etc., having very low conductivity are known as insulators. Substances like silicon, doped silicon and gallium arsenide having conductivity between conductors and insulators are called semiconductors and are important electronic materials. Certain materials called superconductors by definition have zero resistivity or infinite conductivity. Earlier, only metals and their alloys at very low temperatures (0 to 15 K) were known to behave as superconductors, but nowadays a number of ceramic materials and mixed oxides are also known to show superconductivity at temperatures as high as 150 K.

Electrical conductance through metals is called metallic or electronic conductance and is due to the movement of electrons. The electronic conductance depends on

(i) the nature and structure of the metal
(ii) the number of valence electrons per atom
(iii) temperature (it decreases with increase of temperature).

---

* *Electronically conducting polymers – In 1977 MacDiarmid, Heeger and Shirakawa discovered that acetylene gas can be polymerised to produce a polymer, polyacetylene when exposed to vapours of iodine acquires metallic lustre and conductivity. Since then several organic conducting polymers have been made such as polyaniline, polypyrrole and polythiophene. These organic metals, being composed wholly of elements like carbon, hydrogen and occasionally nitrogen, oxygen or sulphur, are much lighter than normal metals and can be used for making light-weight batteries. Besides, they have the mechanical properties of polymers such as flexibility so that one can make electronic devices such as transistors that can bend like a sheet of plastic. For the discovery of conducting polymers, MacDiarmid, Heeger and Shirakawa were awarded the Nobel Prize in Chemistry for the year 2000.*

As the electrons enter at one end and go out through the other end, the composition of the metallic conductor remains unchanged. The mechanism of conductance through semiconductors is more complex.

We already know (Class XI, Unit 7) that even very pure water has small amounts of hydrogen and hydroxyl ions (~10^{-7}M) which lend it very low conductivity (3.5×10^{-5} S m^{-1}). When electrolytes are dissolved in water, they furnish their own ions in the solution hence its conductivity also increases. The conductance of electricity by ions present in the solutions is called electrolytic or ionic conductance. The conductivity of electrolytic (ionic) solutions depends on:

(i) the nature of the electrolyte added
(ii) size of the ions produced and their solvation
(iii) the nature of the solvent and its viscosity
(iv) concentration of the electrolyte
(v) temperature (it increases with the increase of temperature).

Passage of direct current through ionic solution over a prolonged period can lead to change in its composition due to electrochemical reactions (Section 3.4.1).

Measurement of the Conductivity of Ionic Solutions

We know that accurate measurement of an unknown resistance can be performed on a Wheatstone bridge. However, for measuring the resistance of an ionic solution we face two problems. Firstly, passing direct current (DC) changes the composition of the solution. Secondly, a solution cannot be connected to the bridge like a metallic wire or other solid conductor. The first difficulty is resolved by using an alternating current (AC) source of power. The second problem is solved by using a specially designed vessel called **conductivity cell.** It is available in several designs and two simple ones are shown in the below Fig.3.4.

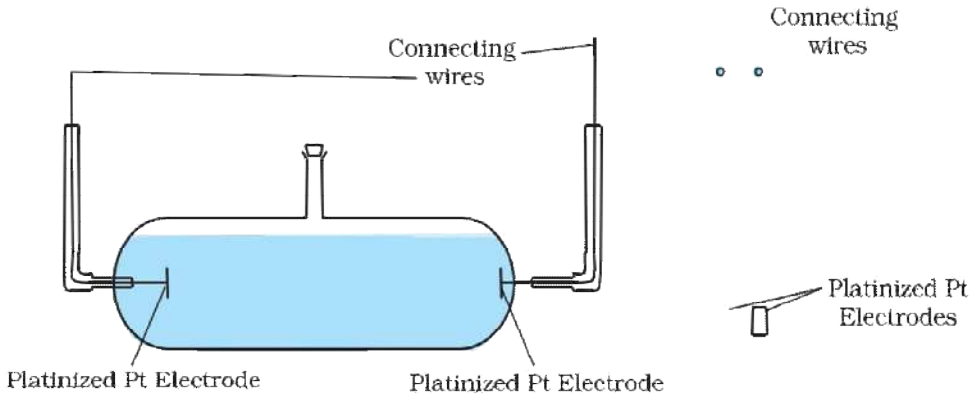


Fig. 3.4 Two different types of conductivity cells.

Basically it consists of two platinum electrodes coated with platinum black (finely divided metallic Pt is deposited on the electrodes electrochemically). These have area of cross section equal to '*A*' and are separated by distance '*l*'. Therefore, solution confined between these electrodes is a column of length *l* and area of cross section *A*. The resistance of such a column of solution is then given by the equation:

$$R = \rho \frac{l}{A} = \frac{l}{\kappa A}$$

The quantity l/A is called cell constant denoted by the symbol, G^*. It depends on the distance between the electrodes and their area of cross-section and has the dimension of length^{-1} and can be calculated if we know l and A. Measurement of l and A is not only inconvenient but also unreliable. The cell constant is usually determined by measuring the resistance of the cell containing a solution whose conductivity is already known. For this purpose, we generally use KCl solutions whose conductivity is known accurately at various concentrations and at different temperatures in Table 32. The cell constant, G^*, is then given by the equation:

$$G^* = \frac{l}{A} = R\,K$$

Table 3.2 Conductivity and Molar conductivity of KCl solutions at 298.15K

Molarity		Concentration Conductivity		Molar Conductivity	
mol L^{-1}	mol m^{-3}	S cm^{-1}	S m^{-1}	S cm^2 mol^{-1}	S m^2 mol^{-1}
1.000	1000	0. 1113	11.13	111.3	111.3×10^{-4}
0.100	100.0	0.0129	1.29	129.0	129.0×10^{-4}
0.010	10.00	0. 00141	0.141	141.0	141.0×10^{-4}

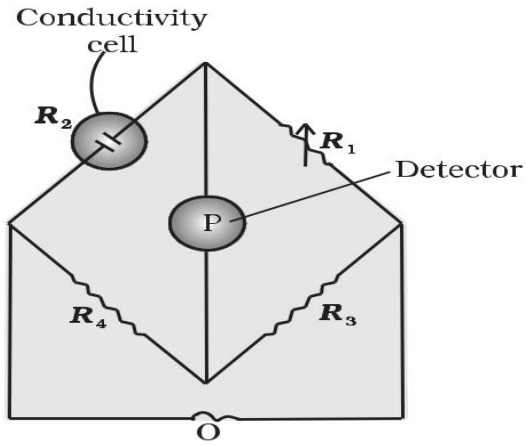


Fig.3.5 Arrangement for measurement of resistance of a solution of an electrolyte.

Once the cell constant is determined, we can use it for measuring the resistance or conductivity of any solution. The set up for the measurement of the resistance is shown in the above Fig.3.5

It consists of two resistances R_3 and R_4, a variable resistance $R1$ and the conductivity cell having the unknown resistance R_2. The Wheatstone bridge is fed by an oscillator O (a source of a.c. power in the audio frequency range 550 to 5000 cycles per second). P is a suitable detector (a headphone or other electronic device) and the bridge is balanced when no current passes through the detector. Under these conditions:

$$\text{Unknown resistance } R_2 = \frac{R_1\,R_4}{R_3}$$

These days, inexpensive conductivity meters are available which can directly read the conductance or resistance of the solution in the conductivity cell. Once the cell constant and

the resistance of the solution in the cell are determined, the conductivity of the solution is given by the equation:

$$\kappa = \frac{\text{cell constant}}{R} = \frac{G^*}{R}$$

The conductivity of solutions of different electrolytes in the same solvent and at a given temperature differs due to charge and size of the ions in which they dissociate, the concentration of ions or ease with which the ions move under a potential gradient. It, therefore, becomes necessary to define a physically more meaningful quantity called **molar conductivity** denoted by the symbol Λm (Greek, lambda). It is related to the conductivity of the solution by the equation:

$$\text{Molar conductivity} = \Lambda_m = \frac{\kappa}{c}$$

In the above equation, if κ is expressed in S m^{-1} and the concentration,

c in mol m^{-3} then the units of Λm are in S m^2 mol^{-1}. It may be noted that:
1 mol m^{-3} = 1000(L/m^3) × molarity (mol/L), and hence

$$\Lambda_m(\text{S m}^2\ \text{mol}^{-1}) = \frac{\kappa\ (\text{S m}^{-1})}{1000\ \text{L m}^{-3} \times \text{molarity (mol L}^{-1})}$$

If we use S cm^{-1} as the units for κ and mol cm^{-3}, the units of concentration, then the units for *Ëm* are S cm^2 mol^{-1}. It can be calculated by using the equation:

$$\Lambda_m\,(\text{S cm}^2\ \text{mol}^{-1}) = \frac{\kappa\ (\text{S cm}^{-1}) \times 1000\,(\text{cm}^3/\text{L})}{\text{molarity (mol/L)}}$$

Both type of units are used in literature and are related to each other by the equations:

1 S m^2mol^{-1} = 10^4 S cm^2mol^{-1} or
1 S cm^2mol^{-1} =10^{-4} S m^2mol^{-1}.

Variation of Conductivity and Molar Conductivity with Concentration

Both conductivity and molar conductivity change with the concentration of the electrolyte. Conductivity always decreases with decrease in concentration both, for weak and strong electrolytes. This can be explained by the fact that the number of ions per unit volume that carry the current in a solution decreases on dilution. The conductivity of a solution at any given concentration is the conductance of one unit volume of solution kept between two platinum electrodes with unit area of cross section and at a distance of unit length. This is clear from the equation:

$$G = \frac{\kappa A}{l} = \kappa$$ (both *A* and *l* are unity in their appropriate units in m or cm)

Molar conductivity of a solution at a given concentration is the conductance of the volume *V* of solution containing one mole of electrolyte kept between two electrodes with area of cross section *A* and distance of unit length. Therefore,

$$\Lambda_m = \frac{\kappa A}{l} = \kappa$$

Since $l = 1$ and $A = V$ (volume containing 1 mole of electrolyte)

$$\Lambda_m = \kappa V \qquad (3.22)$$

Molar conductivity increases with decrease in concentration. This is because the total volume, *V*, of solution containing one mole of electrolyte also increases. It has been found that decrease in κ on dilution of a solution is more than compensated by increase in its volume. Physically, it means that at a given concentration, Λ*m* can be defined as the conductance of the electrolytic solution kept between the electrodes of a conductivity cell at unit distance but having area of cross section large enough to accommodate sufficient volume of solution that contains one mole of the electrolyte. When concentration approaches zero, the molar conductivity is known as **limiting molar conductivity** and is represented by the symbol *Ëm*°. The variation in Λ*m* with concentration is different for strong and weak electrolytes.

Strong Electrolytes

For strong electrolytes, Λ increases slowly with dilution and can be represented by the equation:

$$\Lambda_m = \ddot{E}_m^\circ - A\, c^{½}$$

It can be seen that if we plot Λ*m* against *c*1/2, we obtain a straight line with intercept equal to*Ëm*° and slope equal to '–*A*'. The value of the constant '*A*' for a given solvent and temperature depends on the type of electrolyte i.e., the charges on the cation and anion produced on the dissociation of the electrolyte in the solution (Fig.3.6). Thus, NaCl, $CaCl_2$, $MgSO_4$ are known as 1-1, 2-1 and 2- 2 electrolytes respectively. All electrolytes of a particular type have the same value for '*A*'.

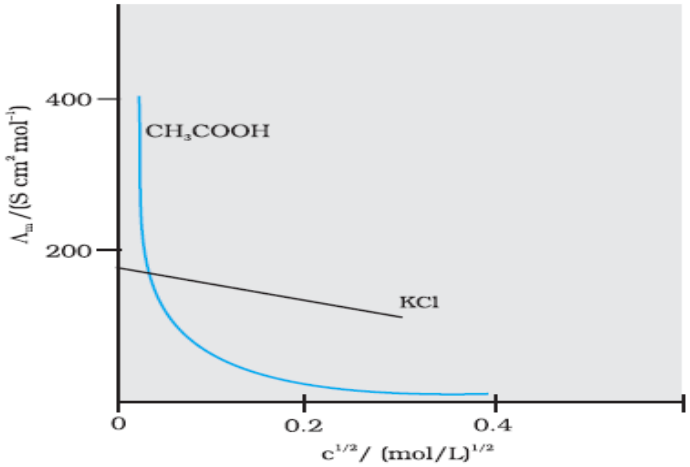


Fig.3.6 Molar conductivity versus c½ for acetic acid (weak electrolyte) and potassium chloride (strong electrolyte) in aqueous solutions.

Kohlrausch examined *Ëm*° values for a number of strong electrolytes and observed certain regularities. He noted that the difference in *Ëm*° of the electrolytes NaX and KX for any X is nearly constant. For example at 298 K:

$$\ddot{E}^{\circ}_{m\,(KCl)} - \ddot{E}^{\circ}_{m\,(NaCl)} = \ddot{E}^{\circ}_{m\,(KBr)} - \ddot{E}^{\circ}_{m\,(NaBr)}$$
$$= \ddot{E}^{\circ}_{m\,(KI)} - \ddot{E}^{\circ}_{m\,(NaI)} \simeq 23.4 \text{ S cm}^2 \text{ mol}^{-1}$$

and similarly it was found that

$$\ddot{E}^{\circ}_{m\,(NaBr)} - \ddot{E}^{\circ}_{m\,(NaCl)} = \ddot{E}^{\circ}_{m\,(KBr)} - \ddot{E}^{\circ}_{m\,(KCl)} \simeq 1.8 \text{ S cm}^2 \text{ mol}^{-1}$$

On the basis of the above observations he enunciated **Kohlrausch law of independent migration of ions**. The law states that *limiting molar conductivity of an electrolyte can be represented as the sum of the individual contributions of the anion and cation of the electrolyte. Thus, if λ°Na+ and λ°Cl – are limiting molar conductivity* of the sodium and chloride ions respectively, then the limiting molar conductivity for sodium chloride is given by the equation:

$$\ddot{E}^{\circ}_{m\,(NaCl)} = \lambda^{0}_{Na^+} + \lambda^{0}_{Cl^-}$$

In general, if an electrolyte on dissociation gives *v*+ cations and *v*– anions then its limiting molar conductivity is given by:

$$\ddot{E}^{\circ}_{m} = \nu_{+}\,\lambda^{0}_{+} + \nu_{-}\,\lambda^{0}_{-}$$

Here,λ +0 andλ −0 are the limiting molar conductivities of the cation and anion respectively. The values of λ0 for some cations and anions at 298 K are given in below Table.3.3.

Table 3.3 Limiting molar conductivity for some ions in water at 298 K

Ion	λ^0/(S cm^{2}mol^{-1})	Ion	λ^0/(S cm^2 mol^{-1})
H^+	349.6	OH^-	199.1
Na^+	50.1	Cl^-	76.3
K^+	73.5	Br^-	78.1
Ca^{2+}	119.0	CH_3COO^-	40.9
Mg^{2+}	106.0	SO_4^{2-}	160.0

Weak electrolytes

Weak electrolytes like acetic acid have lower degree of dissociation at higher concentrations and hence for such electrolytes, the change in Λ*m* with dilution is due to increase in the degree of dissociation and consequently the number of ions in total volume of solution that contains 1 mol of electrolyte. In such cases *Ëm* increases steeply on dilution, especially near lower concentrations. Therefore, *Ëm*° cannot be obtained by extrapolation of Λ*m* to zero concentration. At infinite dilution (i.e., concentration c → zero) electrolyte dissociates completely (α =1), but at such low concentration the conductivity of the solution is so low that it cannot be measured accurately. Therefore, *Ëm* ° for weak electrolytes is obtained by using Kohlrausch law of independent migration of ions (Example 3.8). At any concentration c, if α is the degree of dissociation then it can be approximated to the ratio of molar conductivity *Ëm* at the concentration c to limiting molar conductivity, *Ëm* °. Thus we have:

$$\alpha = \frac{\Lambda_m}{\Lambda_m^\circ}$$

But we know that for a weak electrolyte like acetic acid (Class XI, Unit 7),

$$K_a = \frac{c\alpha^2}{(1-\alpha)} = \frac{c\Lambda_m^2}{\Lambda_m^{\circ 2}\left(1-\frac{\Lambda_m}{\Lambda_m^\circ}\right)} = \frac{c\Lambda_m^2}{\Lambda_m^\circ\left(\Lambda_m^\circ - \Lambda_m\right)}$$

Applications of Kohlrausch law

Using Kohlrausch law of independent migration of ions, it is possible to calculate *Ëm*° for any electrolyte from the λo of individual ions. Moreover, for weak electrolytes like acetic acid it is possible to determine the value of its dissociation constant once we know the *Ëm* ° and Λ*m* at a given concentration c.

3.5 Electrolytic Cells and Electrolysis

In an **electrolytic cell** external source of voltage is used to bring about a chemical reaction. The electrochemical processes are of great importance in the laboratory and the chemical industry. One of the simplest electrolytic cell consists of two copper strips dipping in an aqueous solution of copper sulphate. If a DC voltage is applied to the two electrodes, then Cu 2^+ ions discharge at the cathode (negatively charged) and the following reaction takes place:

$$Cu^{2+}(aq) + 2e^- \rightarrow Cu\ (s)$$

Copper metal is deposited on the cathode. At the anode, copper is converted into $Cu2^+$ ions by the reaction:

$$Cu(s) \rightarrow Cu^{2+}(s) + 2e^-$$

Thus copper is dissolved (oxidised) at anode and deposited (reduced) at cathode. This is the basis for an industrial process in which impure copper is converted into copper of high purity. The impure copper is made an anode that dissolves on passing current and pure copper is deposited at the cathode. Many metals like Na, Mg, Al, etc. are produced on large scale by electrochemical reduction of their respective cations where no suitable chemical reducing agents are available for this purpose.

Sodium and magnesium metals are produced by the electrolysis of their fused chlorides and aluminium is produced (Class XII, Unit 6) by electrolysis of aluminium oxide in presence of cryolite.

Quantitative Aspects of Electrolysis

Michael Faraday was the first scientist who described the quantitative aspects of electrolysis. Now Faraday's laws also flow from what has been discussed earlier.

Faraday's Laws of Electrolysis

After his extensive investigations on electrolysis of solutions and melts of electrolytes, Faraday published his results during 1833-34 in the form of the following well known Faraday's two laws of electrolysis:

1. First Law

The amount of chemical reaction which occurs at any electrode during electrolysis by a current is proportional to the quantity of electricity passed through the electrolyte (solution or melt).

2. Second Law

The amounts of different substances liberated by the same quantity of electricity passing through the electrolytic solution are proportional to their chemical equivalent weights (Atomic Mass of Metal ÷ Number of electrons required to reduce the cation).

There were no constant current sources available during Faraday's times. The general practice was to put a coulometer (a standard electrolytic cell) for determining the quantity of electricity passed from the amount of metal (generally silver or copper) deposited or consumed. However, coulometers are now obsolete and we now have constant current (*I*) sources available and the quantity of electricity *Q*, passed is given by

$$Q = It$$

Q is in coloumbs when *I* is in ampere and *t* is in second.

The amount of electricity (or charge) required for oxidation or reduction depends on the stoichiometry of the electrode reaction. For example, in the reaction:

$$Ag^{+}(aq) + e^{-} \rightarrow Ag(s)$$

One mole of the electron is required for the reduction of one mole of silver ions. We know that charge on one electron is equal to 1.6021× 10–$_{19}$C. Therefore, the charge on one mole of electrons is equal to:

$$N_A \times 1.6021 \times 10^{-19} \quad C = 6.02 \times 10^{23}\ mol^{-1} \times 1.6021 \times 10^{-19}$$
$$C = 96487\ C\ mol^{-1}$$

This quantity of electricity is called **Faraday** and is represented by the symbol **F**.
For approximate calculations we use 1F $\simeq$96500 C mol^{-1}. For the electrode reactions:

$$Mg^{2+}(l) + 2e^{-} \longrightarrow Mg(s)$$
$$Al^{3+}(l) + 3e^{-} \longrightarrow Al(s)$$

It is obvious that one mole of Mg^{2+} and Al^{3+} require 2 mol of electrons (2F) and 3 mol of electrons (3F) respectively. The charge passed through the electrolytic cell during electrolysis is equal to the product of current in amperes and time in seconds. In commercial production of metals, current as high as 50,000 amperes are used that amounts to about 0.518 F per second.

Products of Electrolysis

Products of electrolysis depend on the nature of material being electrolysed and the type of electrodes being used. If the electrode is inert (e.g., platinum or gold), it does not participate in the chemical reaction and acts only as source or sink for electrons. On the other hand, if the electrode is reactive, it participates in the electrode reaction. Thus, the products of electrolysis may be different for reactive and inert electrodes.The products of electrolysis depend on the different oxidising and reducing species present in the electrolytic cell and their standard electrode potentials. Moreover, some of the electrochemical processes although feasible, are so slow kinetically that at lower voltages these don't seem to take place and extra potential (called *overpotential*) has to be applied, which makes such process more difficult to occur.

For example, if we use molten NaCl, the products of electrolysis are sodium metal and Cl^2 gas. Here we have only one cation (Na^+) which is reduced at the cathode ($Na^{+\,+}$ e– → Na) and one anion (Cl^-) which is oxidised at the anode (Cl–→ ½Cl2$^+$e–) . During the electrolysis of aqueous sodium chloride solution, the products are NaOH, Cl2 and H2. In this case besides Na^+ and Cl^- ions we also have H^+ and OH^- ions along with the solvent molecules, H_2O.

At the cathode there is competition between the following reduction reactions:

$Na^+ (aq) + e^- \rightarrow Na (s)$ $\quad E^\ominus_{(cell)} = -2.71$ V

$H^+ (aq) + e^- \rightarrow \frac{1}{2} H_2 (g)$ $\quad E^\ominus_{(cell)} = 0.00$ V

The reaction with higher value of $E\_$ is preferred and, therefore, the reaction at the cathode during electrolysis is:

$H^+ (aq) + e^- \rightarrow \frac{1}{2} H_2 (g)$ (3.33)

but $H^+ (aq)$ is produced by the dissociation of H_2O, i.e.,

$H_2O (l) \rightarrow H^+ (aq) + OH^- (aq)$ (3.34)

Therefore, the net reaction at the cathode may be written as the sum of (3.33) and (3.34) and we have

$H_2O (l) + e^- \rightarrow \frac{1}{2}H_2(g) + OH^-$ (3.35)

At the anode the following oxidation reactions are possible:

$Cl^- (aq) \rightarrow \frac{1}{2} Cl_2 (g) + e^-$ $\quad E^\ominus_{(cell)} = 1.36$ V (3.36)

$2H_2O (l) \rightarrow O_2 (g) + 4H^+(aq) + 4e^-$ $\quad E^\ominus_{(cell)} = 1.23$ V (3.37)

The reaction at anode with lower value of $E\_$ is preferred and therefore, water should get oxidised in preference to Cl– (aq). However, on account of overpotential of oxygen, reaction (3.36) is preferred. Thus, the net reactions may be summarised as:

$NaCl (aq) \xrightarrow{H_2O} Na^+ (aq) + Cl^- (aq)$

Cathode: $H_2O(l) + e^- \rightarrow \frac{1}{2} H_2(g) + OH^- (aq)$

Anode: $Cl^- (aq) \rightarrow \frac{1}{2} Cl_2(g) + e^-$

Net reaction:

$NaCl(aq) + H_2O(l) \rightarrow Na^+(aq) + OH^-(aq) + \frac{1}{2}H_2(g) + \frac{1}{2}Cl_2(g)$

The standard electrode potentials are replaced by electrode potentials given by Nernst equation (Eq. 3.8) to take into account the concentration effects. During the electrolysis of sulphuric acid, the following processes are possible at the anode:

$2H_2O(l) \rightarrow O_2(g) + 4H^+(aq) + 4e^-$ $\quad E^\ominus_{(cell)} = +1.23$ V, (3.38)

$2SO_4^{2-} (aq) \rightarrow S_2O_8^{2-} (aq) + 2e^-$ $\quad E^\ominus_{(cell)} = 1.96$ V (3.39)

For dilute sulphuric acid, reaction (3.38) is preferred but at higher concentrations of H_2SO_4 process, reaction (3.39) is preferred.

3.6 Batteries

Any battery (actually it may have one or more than one cell connected in series) or cell that we use as a source of electrical energy is basically a galvanic cell where the chemical energy of the redox reaction is converted into electrical energy. However, for a battery to be of practical use it should be reasonably light, compact and its voltage should not vary appreciably during its use. There are mainly two types of batteries.

3.6.1 Primary Batteries

In the primary batteries, the reaction occurs only once and after use over a period of time battery becomes dead and cannot be reused again. The most familiar example of this type is the dry cell (known as Leclanche cell after its discoverer) which is used commonly in our transistors and clocks. The cell consists of a zinc container that also acts as anode and the cathode is a carbon (graphite) rod surrounded by powdered manganese dioxide and carbon (Fig.3.7). The space between the electrodes is filled by a moist paste of ammonium chloride (NH_4Cl) and zinc chloride ($ZnCl_2$). The electrode reactions are complex, but they can be written approximately as follows:

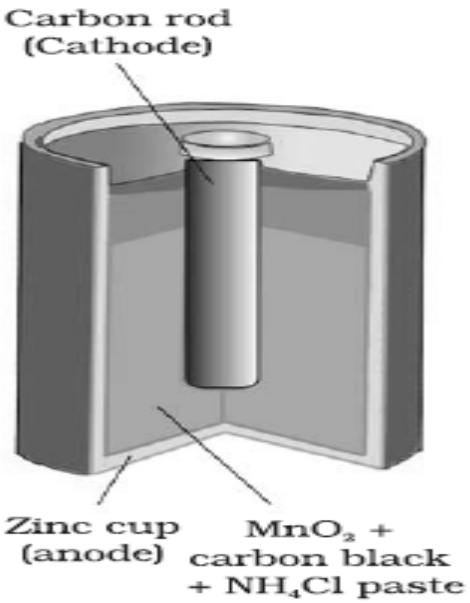


Fig.3.7 A commercial dry cell consists of a graphite (carbon) cathode in a zinc container; the latter acts as the anode.

Anode: $Zn(s) \longrightarrow Zn^{2+} + 2e^-$

Cathode: $MnO_2 + NH_4^+ + e^- \longrightarrow MnO(OH) + NH_3$

In the reaction at cathode, manganese is reduced from the + 4 oxidation state to the +3 state. Ammonia produced in the reaction forms a complex with Zn^{2+} to give $[Zn(NH_3)_4]^{2+}$. The cell has a potential of nearly 1.5 V. Mercury cell, (Fig. 3.8) suitable for low current devices like hearing aids, watches, etc. consists of zinc – mercury amalgam as anode and a paste of HgO and carbon as the cathode. The electrolyte is a paste of KOH and ZnO. The electrode reactions for the cell are given below:

Anode: $Zn(Hg) + 2OH^- \longrightarrow ZnO(s) + H_2O + 2e^-$

Cathode: $HgO + H_2O + 2e^- \longrightarrow Hg(l) + 2OH^-$

The overall reaction is represented by

$$Zn(Hg) + HgO(s) \longrightarrow ZnO(s) + Hg(l)$$

The cell potential is approximately 1.35 V and remains constant during its life as the overall reaction does not involve any ion in solution whose concentration can change during its life time.

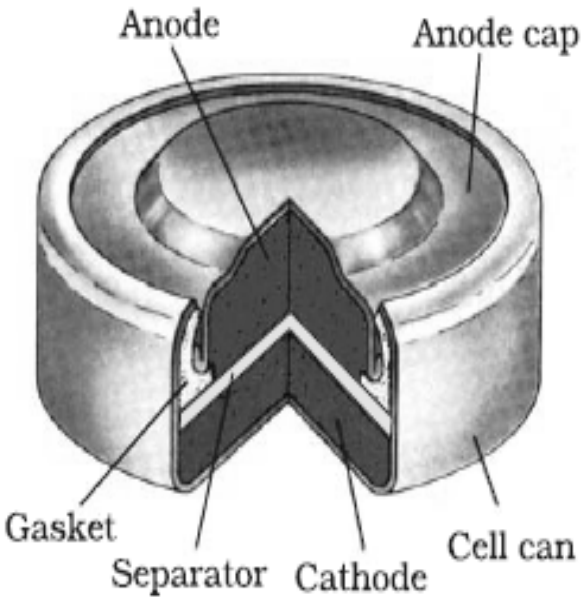


Fig 3.8 Commonly used mercury cell. The reducing agent is zinc and the oxidising agent is mercury (II) oxide.

Secondary Batteries

Secondary cell after use can be recharged by passing current through it in the opposite direction so that it can be used again. A good secondary cell can undergo a large number of discharging and charging cycles. The most important secondary cell is the lead storage battery (Fig.3.9) commonly used in automobiles and invertors. It consists of a lead anode and a grid of lead packed with lead dioxide (PbO_2) as cathode. A 38% solution of sulphuric acid is used as an electrolyte.

The cell reactions when the battery is in use are given below:

$$\text{Anode: } Pb(s) + SO_4^{2-}(aq) \rightarrow PbSO_4(s) + 2e^-$$

$$\text{Cathode: } PbO_2(s) + SO_4^{2-}(aq) + 4H^+(aq) + 2e^- \rightarrow PbSO_4(s) + 2H_2O(l)$$

i.e., overall cell reaction consisting of cathode and anode reactions is:

$$Pb(s) + PbO_2(s) + 2H_2SO_4(aq) \rightarrow 2PbSO_4(s) + 2H_2O(l)$$

On charging the battery the reaction is reversed and $PbSO_4(s)$ on anode and cathode is converted into Pb and PbO_2, respectively.

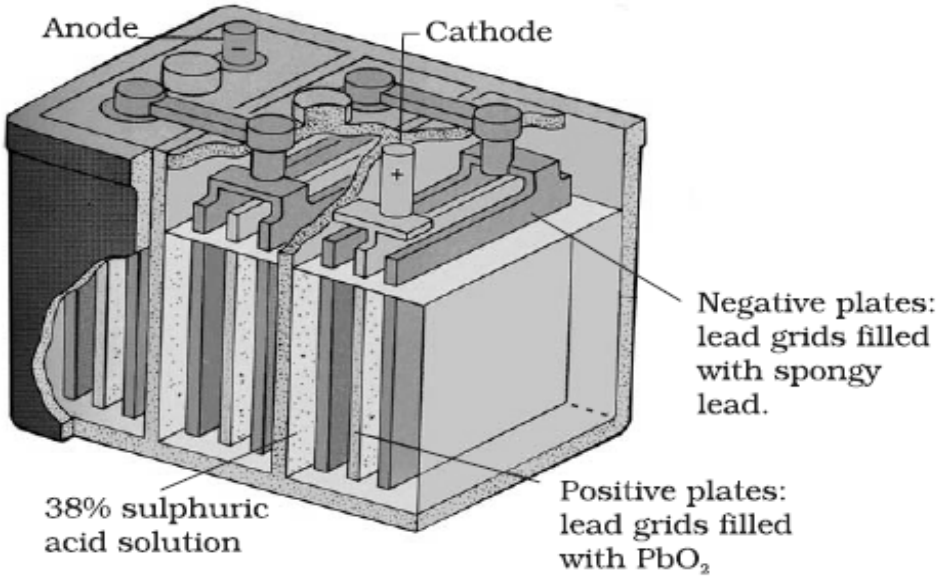


Fig.3.9 The Lead storage battery

Another important secondary cell is the nickel cadmium cell, which has longer life than the lead storage cell (Fig.3.10) but more expensive to manufacture. We shall not go into details of working of the cell and the electrode reactions during charging and discharging. The overall reaction during discharge is:

$$Cd\ (s) + 2Ni(OH)_3\ (s) \rightarrow CdO\ (s) + 2Ni(OH)_2\ (s) + H_2O(l)$$

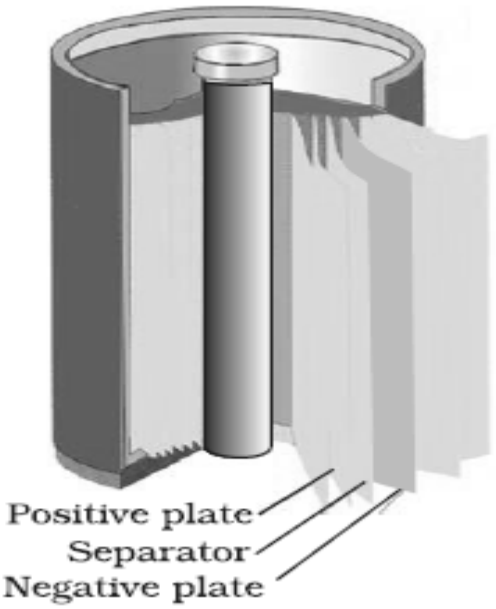


Fig.3.10 A rechargeable nickel-cadmium cell in a jelly roll arrangement and separated by a layer soaked in moist sodium or potassium hydroxide.

3.7 Fuel Cells

Production of electricity by thermal plants is not a very efficient method and is a major source of pollution. In such plants, the chemical energy (heat of combustion) of fossil fuels (coal, gas or oil) is first used for converting water into high pressure steam. This is then used to run a turbine to produce electricity. We know that a galvanic cell directly converts chemical energy into electricity and is highly efficient. It is now possible to make such cells in which reactants are fed continuously to the electrodes and products are removed continuously from the electrolyte compartment. Galvanic cells that are designed to convert the energy of combustion of fuels like hydrogen, methane, methanol, etc. directly into electrical energy are called **fuel cells**. One of the most successful fuel cells uses the reaction of hydrogen with oxygen to form water. The cell was used for providing electrical power in the Apollo space programme. The water vapours produced during the reaction were condensed and added to the drinking water supply for the astronauts. In the cell, hydrogen and oxygen are bubbled through porous carbon electrodes into concentrated aqueous sodium hydroxide solution. Catalysts like finely divided platinum or palladium metal are incorporated into the electrodes for increasing the rate of electrode reactions. The electrode reactions are given below Fig.3.11.

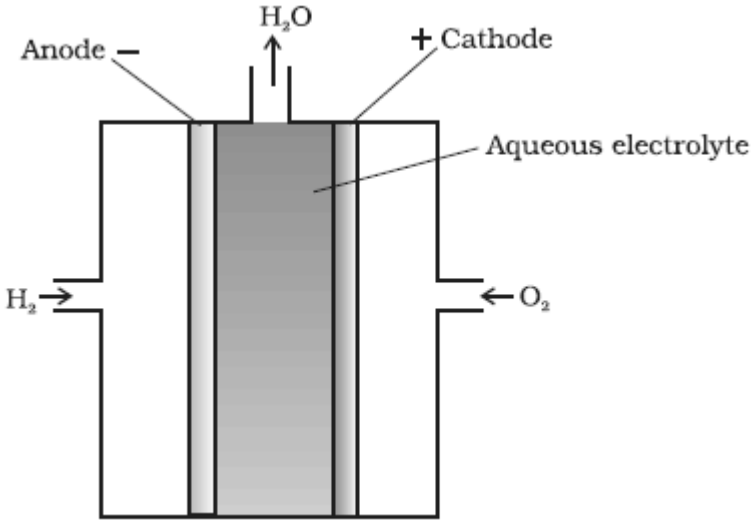


Fig. 3.11 Fuel cell using H2 and O2 produces electricity.

Cathode: $O_2(g) + 2H_2O(l) + 4e^- \longrightarrow 4OH^-(aq)$

Anode: $2H_2(g) + 4OH^-(aq) \longrightarrow 4H_2O(l) + 4e^-$

Overall reaction being:

$2H_2(g) + O_2(g) \longrightarrow 2\ H_2O(l)$

The cell runs continuously as long as the reactants are supplied. Fuel cells produce electricity with an efficiency of about 70 % compared to thermal plants whose efficiency is about 40%. There has been tremendous progress in the development of new electrode materials, better catalysts and electrolytes for increasing the efficiency of fuel cells. These have been used in automobiles on an experimental basis. Fuel cells are pollution free and in view of their future importance, a variety of fuel cells have been fabricated and tried.

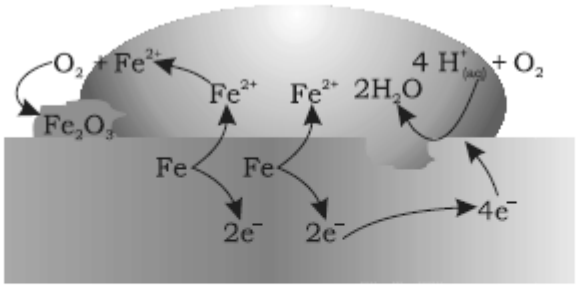
Corrosion

Corrosion slowly coats the surfaces of metallic objects with oxides or other salts of the metal. The rusting of iron, tarnishing of silver, development of green coating on copper and bronze are some of the examples of corrosion. It causes enormous damage to buildings, bridges, ships and to all objects made of metals especially that of iron. We lose crores of rupees every year on account of corrosion.

In corrosion, a metal is oxidised by loss of electrons to oxygen and formation of oxides. Corrosion of iron (commonly known as rusting) occurs in presence of water and air. The chemistry of corrosion is quite complex but it may be considered essentially as an electrochemical phenomenon. At a particular spot (Fig. 3.12) of an object made of iron, oxidation takes place and that spot behaves as anode and we can write the reaction

$$\text{Anode: } 2\ Fe\ (s) \longrightarrow 2\ Fe^{2+} + 4\ e^- \qquad E^{\ominus}_{(Fe^{2+}/Fe)} = -0.44\ V$$

Electrons released at anodic spot move through the metal and go to another spot on the metal and reduce oxygen in presence of H^+ (which is believed to be available from H_2CO_3 formed due to dissolution of carbon dioxide from air into water. Hydrogen ion in water may also be available due to dissolution of other acidic oxides from the atmosphere). This spot behaves as cathode with the reaction



Oxidation: $Fe\ (s) \rightarrow Fe^{2+}\ (aq) + 2e^-$

Reduction: $O_2\ (g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O(l)$

Atomospheric oxidation : $2Fe^{2+}(aq) + 2H_2O(l) + \frac{1}{2}O_2(g) \rightarrow Fe_2O_3(s) + 4H^+(aq)$

Fig. 3.12 Corrosion of iron in atmosphere.

$$\text{Cathode: } O_2(g) + 4\ H^+(aq) + 4\ e^- \longrightarrow 2\ H_2O\ (l) \quad E^{\ominus}_{H^+|O_2|H_2O} = 1.23\ V$$

The overall reaction being:

$$2Fe(s) + O_2(g) + 4H^+(aq) \longrightarrow 2Fe^{2+}(aq) + 2\ H_2O\ (l) \qquad E^{\ominus}_{(cell)} = 1.67\ V$$

The ferrous ions are further oxidised by atmospheric oxygen to ferric ions which come out as rust in the form of hydrated ferric oxide ($Fe_2O_3.\ x\ H_2O$) and with further production of hydrogen ions.

Prevention of corrosion is of prime importance. It not only saves money but also helps in preventing accidents such as a bridge collapse or failure of a key component due to corrosion. One of the simplest methods of preventing corrosion is to prevent the surface of the metallic object to come in contact with atmosphere. This can be done by covering the surface with paint or by some chemicals (e.g. bisphenol). Another simple method is to cover the surface by other metals (Sn, Zn, etc.) that are inert or react to save the object. An electrochemical method is to provide a sacrificial electrode of another metal (like Mg, Zn, etc.) which corrodes itself but saves the object.

The Hydrogen Economy

At present the main source of energy that is driving our economy is fossil fuels such as coal, oil and gas. As more people on the planet aspire to improve their standard of living, their energy requirement will increase. In fact, the per capita consumption of energy used is a measure of development. Of course, it is assumed that energy is used for productive purpose and not merely wasted. We are already aware that carbon dioxide produced by the combustion of fossil fuels is resulting in the 'Greenhouse Effect'. This is leading to a rise in the temperature of the Earth's surface, causing polar ice to melt and ocean levels to rise. This will flood low-lying areas along the coast and some island nations such as Maldives face total submergence. In order to avoid such a catastrope, we need to limit our use of carbonaceous fuels. Hydrogen provides an ideal alternative as its combustion results in water only. Hydrogen production must come from splitting water using solar energy. Therefore, hydrogen can be used as a renewable and non polluting source of energy. This is the vision of the Hydrogen Economy. Both the production of hydrogen by electrolysis of water and hydrogen combustion in a fuel cell will be important in the future. And both these technologies are based on electrochemical principles.

Summary

An **electrochemical cell** consists of two metallic electrodes dipping in electrolytic solution(s). Thus an important component of the electrochemical cell is the ionic conductor or electrolyte. Electrochemical cells are of two types. In **galvanic cell**, the **chemical energy** of a **spontaneous redox reaction** is converted into electrical work, whereas in an electrolytic cell, electrical energy is used to carry out a **non-spontaneous redox reaction**. The **standard electrode potential** for any electrode dipping in an appropriate solution is defined with respect to standard electrode potential of **hydrogen electrode** taken as zero. The standard potential of the cell can be obtained by taking the difference of the standard potentials of cathode and anode $(E^{\ominus}_{(cell)} = E^{\ominus}\text{cathode} - E^{\ominus}\text{anode})$. The standard potential of the cells are related to standard Gibbs energy $(\Delta_r G^{\ominus} = -nF E^{\ominus}_{(cell)})$ and **equilibrium constant** $(\Delta_r G^{\ominus} = -RT \ln K)$ of the reaction taking place in the cell. Concentration dependence of the potentials of the electrodes and the cells are given by Nernst equation.

The **conductivity**, κ, of an electrolytic solution depends on the concentration of the electrolyte, nature of solvent and temperature. **Molar conductivity**, *Ëm*, is defined by = κ /*c* where *c* is the concentration. Conductivity decreases but molar conductivity increases with decrease in concentration. It increases slowly with decrease in concentration for strong electrolytes while the increase is very steep for weak electrolytes in very dilute solutions. Kohlrausch found that molar conductivity at infinite dilution, for an electrolyte is sum of the contribution of the molar conductivity of the ions in which it dissociates. It is known as **law of independent migration of ions** and has many applications. Ions conduct electricity through the solution but oxidation and reduction of the ions take place at the electrodes in an electrochemical cell. **Batteries** and **fuel cells** are very useful forms of galvanic cell. **Corrosion** of metals is essentially an **electrochemical phenomenon**. Electrochemical principles are relevant to the **Hydrogen Economy**.

CHEMICAL KINETICS

Chemistry, by its very nature, is concerned with change. Substances with well defined properties are converted by chemical reactions into other substances with different properties. For any chemical reaction, chemists try to find out

(a) the feasibility of a chemical reaction which can be predicted by thermodynamics (as you know that a reaction with $G < 0$, at constant temperature and pressure is feasible);

* extent to which a reaction will proceed can be determined from chemical equilibrium;
* speed of a reaction i.e. time taken by a reaction to reach equilibrium.

Along with feasibility and extent, it is equally important to know the rate and the factors controlling the rate of a chemical reaction for its complete understanding. For example, which parameters determine as to how rapidly food gets spoiled? How to design a rapidly setting material for dental filling? Or what controls the rate at which fuel burns in an auto engine? All these questions can be answered by the branch of chemistry, which deals with the study of reaction rates and their mechanisms, called **chemical kinetics**. The word kinetics is derived from the Greek word 'kinesis' meaning movement. Thermodynamics tells only about the feasibility of a reaction whereas chemical kinetics tells about the rate of a reaction. For example, thermodynamic data indicate that diamond shall convert to graphite but in reality the conversion rate is so slow that the change is not perceptible at all. Therefore, most people think that diamond is forever. Kinetic studies not only help us to determine the speed or rate of a chemical reaction but also describe the conditions by which the reaction rates can be altered. The factors such as concentration, temperature, pressure and catalyst affect the rate of a reaction. At the macroscopic level, we are interested in amounts reacted or formed and the rates of their consumption or formation. At the molecular level, the reaction mechanisms involving orientation and energy of molecules undergoing collisions are discussed.

In this Unit, we shall be dealing with average and instantaneous rate of reaction and the factors affecting these. Some elementary ideas about the collision theory of reaction rates are also given. However, in order to understand all these, let us first learn about the reaction rate.

3.8 Rate of a Chemical Reaction

Some reactions such as ionic reactions occur very fast, for example, precipitation of silver chloride occurs instantaneously by mixing of aqueous solutions of silver nitrate and sodium chloride. On the other hand, some reactions are very slow, for example, rusting of iron in the presence of air and moisture. Also there are reactions like inversion of cane sugar and hydrolysis of starch, which proceed with a moderate speed. Can you think of more examples from each category?

You must know that speed of an automobile is expressed in terms of change in the position or distance covered by it in a certain period of time. Similarly, the speed of a reaction or the rate of a reaction can be defined as the change in concentration of a reactant or product in unit time. To be more specific, it can be expressed in terms of:

- the rate of decrease in concentration of any one of the reactants, or
- the rate of increase in concentration of any one of the products.

Consider a hypothetical reaction, assuming that the volume of the system remains constant.

$$R \rightarrow P$$

One mole of the reactant R produces one mole of the product P. If $[R]_1$ and $[P]_1$ are the concentrations of R and P respectively at time t_1 and $[R]_2$ and $[P]_2$ are their concentrations at time t_2 then,

$$t = t_2 - t_1$$
$$[R] = [R]_2 - [R]_1$$
$$[P] = [P]_2 - [P]_1$$

The square brackets in the above expressions are used to express molar concentration.

Rate of disappearance of R

$$= \frac{\text{Decrease in concentration of R}}{\text{Time taken}} = -\frac{[R]}{t} \qquad 3.40$$

Rate of appearance of P

$$= \frac{\text{Increase in concentration of P}}{\text{Time taken}} = +\frac{[P]}{t} \qquad 3.41$$

Since, [R] is a negative quantity (as concentration of reactants is decreasing), it is multiplied with –1 to make the rate of the reaction a positive quantity. Equations (3.40) and (3.41) given above represent the average rate of a reaction, r_{av}.

Average rate depends upon the change in concentration of reactants or products and the time taken for that change to occur (Fig. 3.13).

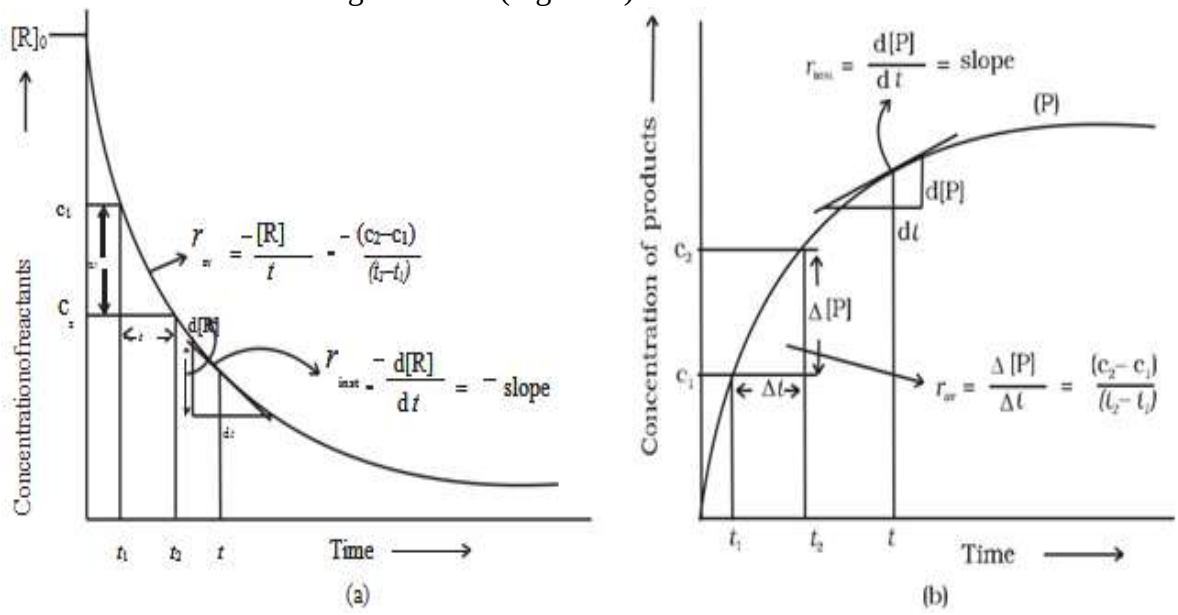


Fig.3.13 Instantaneous and average rate of a reaction

Units of rate of a reaction

From the above equations, it is clear that units of rate are concentration time^{-1}. For example, if concentration is in mol L^{-1} and time is in seconds then the units will be mol $L^{-1}s^{-1}$. However, in gaseous reactions, when the concentration of gases is expressed in terms of their partial pressures, then the units of the rate equation will be atm s^{-1}.

Table 3.4 Average rates of hydrolysis of butyl chloride

$[C_4H_9Cl]_1$ / mol L^{-1}	$[C_4H_9Cl]_2$ / mol L^{-1}	t_1/s	t_2/s	$r_{av} \times 10^4$/mol $L^{-1}s^{-1}$ $= -\{[C_4H_9Cl]_{t_2} - [C_4H_9Cl]_{t_1}\}/(t_2 - t_1) \times 10^4$
0.100	0.0905	0	50	1.90
0.0905	0.0820	50	100	1.70
0.0820	0.0741	100	150	1.58
0.0741	0.0671	150	200	1.40
0.0671	0.0549	200	300	1.22
0.0549	0.0439	300	400	1.10
0.0439	0.0335	400	500	1.04
0.0210	0.017	700	800	0.4

It can be seen from the above Table 3.4 that the average rate falls from 1.90×0^{-4} mol $L^{-1}s^{-1}$ to 0.4×10^{-4} = mol $L^{-1}s^{-1}$. However, average rate cannot be used to predict the rate of a reaction at a particular instant as it would be constant for the time interval for which it is calculated. So, to express the rate at a particular moment of time we determine the **instantaneous rate**. It is obtained when we consider the average rate at the smallest time interval say d*t* (i.e. when *t* approaches zero). Hence, mathematically for an infinitesimally small d*t* instantaneous rate is given by

$$r_{av} = \frac{-[R]}{t} = \frac{[P]}{t}$$

As $t \to 0$ or $r_{inst} = \frac{-d[R]}{dt} = \frac{d[P]}{dt}$

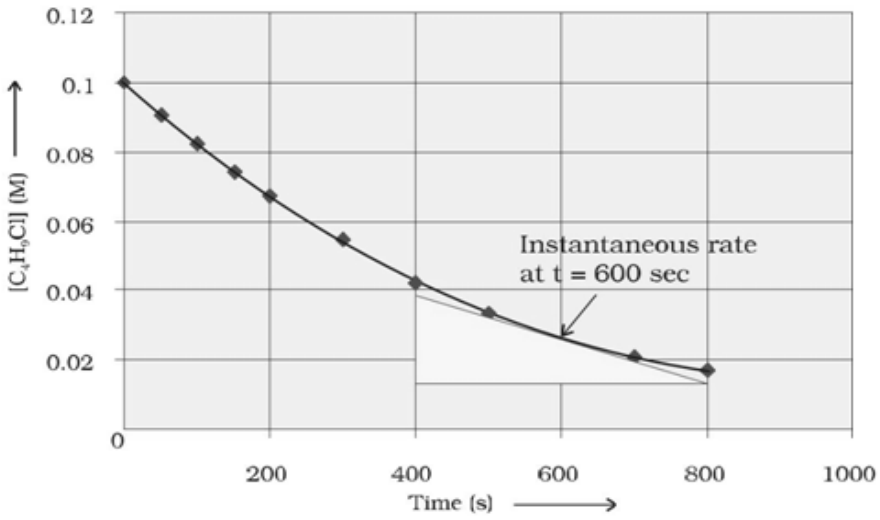


Fig. 3.14 Instantaneous rate of hydrolysis of butyl chloride(C_4H_9Cl)

It can be determined graphically by drawing a tangent at time t on either of the curves for concentration of R and P vs time t and calculating its slope shown in the above Fig.3.14. So in above problem r_{inst} at 600s for example, can be calculated by plotting concentration of butyl chloride as a function of time. A tangent is drawn that touches the curve at *t* = 600s.

The slope of this tangent gives the instantaneous rate.

$$\text{So, } r_{inst} \text{ at } 600\text{ s} = \frac{0.0165 - 0.037}{(800 - 400)\text{ s}}\ \text{mol L}^{-1} = 5.12 \times 10^{-5}\ \text{mol L}^{-1}\text{s}^{-1}$$

At $t = 250$ s $\quad r_{inst} = 1.22 \times 10^{-4}$ mol $L^{-1}s^{-1}$

$t = 350$ s $\quad r_{inst} = 1.0 \times 10^{-4}$ mol $L^{-1}s^{-1}$

$t = 450$ s $\quad r_{inst} = 6.4 \times 10^{-5}$ mol $L^{-1}s^{-1}$

Now consider a reaction

$$Hg(l) + Cl_2(g) \rightarrow HgCl_2(s)$$

Where stoichiometric coefficients of the reactants and products are where stoichiometric coefficients of the reactants and products are

$$\text{Rate of reaction} = -\frac{[Hg]}{t} = -\frac{[Cl_2]}{t} = \frac{[HgCl_2]}{t}$$

i.e., rate of disappearance of any of the reactants is same as the rate of appearance of the products. But in the following reaction, two moles of HI decompose to produce one mole each of H_2 and I_2,

$$2HI(g) \rightarrow H_2(g) + I_2(g)$$

$$\text{Rate of reaction} = -\frac{1}{2}\frac{[HI]}{t} = \frac{[H_2]}{t} = \frac{[I_2]}{t}$$

Similarly, for the reaction

$$5\,Br^-(aq) + BrO_3^-(aq) + 6\,H^+(aq) \rightarrow 3\,Br_2(aq) + 3\,H_2O(l)$$

$$\text{Rate} = -\frac{1}{5}\frac{[Br^-]}{t} = -\frac{[BrO_3^-]}{t} = -\frac{1}{6}\frac{[H^+]}{t} = \frac{1}{3}\frac{[Br_2]}{t} = \frac{1}{3}\frac{[H_2O]}{t}$$

For a gaseous reaction at constant temperature, concentration is directly proportional to the partial pressure of a species and hence, rate can also be expressed as rate of change in partial pressure of the reactant or the product.

3.9 Factors Influencing Rate of a Reaction

Rate of reaction depends upon the experimental conditions such as concentration of reactants (pressure in case of gases), temperature and catalyst.

Dependence of Rate on Concentration

The rate of a chemical reaction at a given temperature may depend on the concentration of one or more reactants and products. The representation of rate of reaction in terms of concentration of the reactants is known as **rate law**. It is also called as rate equation or rate expression.

Rate Expression and Rate Constant

The results in Table 4.1 clearly show that rate of a reaction decreases with the passage of time as the concentration of reactants decrease. Conversely, rates generally increase when reactant concentrations increase. So, rate of a reaction depends upon the concentration of reactants.

Consider a general reaction

$$aA + bB \rightarrow cC + dD$$

where a, b, c and d are the stoichiometric coefficients of reactants and products.
The rate expression for this reaction is

$$\text{Rate} \propto [A]^x [B]^y$$

where exponents x and y may or may not be equal to the stoichiometric coefficients (a and b) of the reactants. Above equation can also be written as

$$\text{Rate} = k[A]^x [B]^y$$

$$-\frac{d[R]}{dt} = k[A]^x [B]^Y$$

This form of equation is known as differential rate equation, where *k* is a proportionality constant called **rate constant**. The equation like, which relates the rate of a reaction to concentration of reactants, is called rate law or rate expression. Thus, **rate law is the expression in which reaction rate is given in terms of molar concentration of reactants with each term raised to some power, which may or may not be same as the stoichiometric coefficient of the reacting species in a balanced chemical equation**. For example:

$$2NO(g) + O_2(g) \rightarrow 2NO_2(g)$$

We can measure the rate of this reaction as a function of initial concentrations either by keeping the concentration of one of the reactants constant and changing the concentration of the other reactant or by changing the concentration of both the reactants. The following results are obtained in Table 3.5.

Table 3.5 Initial rate of formation of NO_2

Experiment	Initial [NO]/ mol L^{-1}	Initial [O_2]/ mol L^{-1}	Initial rate of formation of NO_2 / mol $L^{-1}s^{-1}$
1.	0.30	0.30	0.096
2.	0.60	0.30	0.384
3.	0.30	0.60	0.192
4.	0.60	0.60	0.768

It is obvious, after looking at the results, that when the concentration of NO is doubled and that of O_2 is kept constant then the initial rate increases by a factor of four from 0.096 to 0.384 $molL^{-1}s^{-1}$. This indicates that the rate depends upon the square of the concentration of NO. When concentration of NO is kept constant and concentration of O_2 is doubled the rate also gets doubled indicating that rate depends on concentration of O_2 to the first power. Hence, the rate equation for this reaction will be

$$\text{Rate} = k[NO]^2[O_2]$$

The differential form of this rate expression is given as

$$-\frac{d[R]}{dt} = k[NO]^2[O_2]$$

Now, we observe that for this reaction in the rate equation derived from the experimental data, the exponents of the concentration terms are the same as their stoichiometric coefficients in the balanced chemical equation.

Some other examples are given below:

Reaction	Experimental rate expression
1. $CHCl_3 + Cl_2 \rightarrow CCl_4 + HCl$	Rate = $k[CHCl_3][Cl_2]^{1/2}$
2. $CH_3COOC_2H_5 + H_2O \rightarrow CH_3COOH + C_2H_5OH$	Rate = $k[CH_3COOC_2H_5]^1[H_2O]^0$

In these reactions, the exponents of the concentration terms are not the same as their stoichiometric coefficients. Thus, we can say that:

Rate law for any reaction cannot be predicted by merely looking at the balanced chemical equation, i.e., theoretically but must be determined experimentally.

Order of a Reaction

In the rate equation

$$\text{Rate} = k[A]^x [B]^y$$

x and y indicate how sensitive the rate is to the change in concentration of A and B. Sum of these exponents, i.e., x + y in (4.4) gives the overall order of a reaction whereas x and y represent the order with respect to the reactants A and B respectively.

Hence, **the sum of powers of the concentration of the reactants in the rate law expression is called the order of that chemical reaction**.

Order of a reaction can be 0, 1, 2, 3 and even a fraction. A zero order reaction means that the rate of reaction is independent of the concentration of reactants.

A balanced chemical equation never gives us a true picture of how a reaction takes place since rarely a reaction gets completed in one step. The reactions taking place in one step are called **elementary reactions**. When a sequence of elementary reactions (called mechanism) gives us the products, the reactions are called **complex reactions**.

These may be consecutive reactions (e.g., oxidation of ethane to CO_2 and H_2O passes through a series of intermediate steps in which alcohol, aldehyde and acid are formed), reverse reactions and side reactions (e.g., nitration of phenol yields *o*-nitrophenol and *p*-nitrophenol).

For a general reaction

$$aA + bB \rightarrow cC + dD$$

$$\text{Rate} = k[A]^x [B]^y$$

Where x + y = n = order of the reaction

$$k = \frac{\text{Rate}}{[A]^x [B]^y}$$

$$= \frac{\text{concentration}}{\text{time}} \times \frac{1}{(\text{concerntration})^n}$$

Taking SI units of concentration, mol L^{-1} and time, s, the units of *k* for different reaction order are listed in the below mentioned Table 3.6.

Table 3.6 Units of rate constant

Reaction	Order	Units of rate constant
Zero order reaction	0	$\frac{\text{mol L}^{-1}}{\text{s}} \times \frac{1}{(\text{mol L}^{-1})^0} = \text{mol L}^{-1}\text{ s}^{-1}$
First order reaction	1	$\frac{\text{mol L}^{-1}}{\text{s}} \times \frac{1}{(\text{mol L}^{-1})^1} = \text{s}^{-1}$
Second order reaction	2	$\frac{\text{mol L}^{-1}}{\text{s}} \times \frac{1}{(\text{mol L}^{-1})^2} = \text{mol}^{-1}\text{ L s}^{-1}$

Molecularity of a Reaction

Another property of a reaction called molecularity helps in understanding its mechanism. **The number of reacting species (atoms, ions or molecules) taking part in an elementary reaction, which must collide simultaneously in order to bring about a chemical reaction is called molecularity of a reaction**. The reaction can be uni-molecular when one reacting species is involved, for example, decomposition of ammonium nitrite.

$$NH_4NO_2 \rightarrow N_2 + 2H_2O$$

Bimolecular reactions involve simultaneous collision between two species, for example, dissociation of hydrogen iodide.

$$2HI \rightarrow H_2 + I_2$$

Trimolecular or termolecular reactions involve simultaneous collision between three reacting species, for example,

$$2NO + O_2 \rightarrow 2NO_2$$

The probability that more than three molecules can collide and react simultaneously is very small. Hence, the molecularity greater than three is not observed.

It is, therefore, evident that complex reactions involving more than three molecules in the stoichiometric equation must take place in more than one step.

$$KClO_3 + 6FeSO_4 + 3H_2SO_4 \rightarrow KCl + 3Fe_2(SO_4)_3 + 3H_2O$$

This reaction which apparently seems to be of tenth order is actually a second order reaction. This shows that this reaction takes place in several steps. Which step controls the rate of the overall reaction? The question can be answered if we go through the mechanism of reaction, for example, chances to win the relay race competition by a team depend upon the slowest person in the team. Similarly, the overall rate of the reaction is controlled by the slowest step in a reaction called the **rate determining step**. Consider the decomposition of hydrogen peroxide which is catalysed by iodide ion in an alkaline medium.

$$2H_2O_2 \xrightarrow[\text{Alkaline medium}]{I^-} 2H_2O + O_2$$

The rate equation for this reaction is found to be

$$\text{Rate} = \frac{-d[H_2O_2]}{dt} = k[H_2O_2][I^-]$$

This reaction is first order with respect to both H_2O_2 and I^-. Evidences suggest that this reaction takes place in two steps

(1) $H_2O_2 + I^- \rightarrow H_2O + IO^-$

(2) $H_2O_2 + IO^- \rightarrow H_2O + I^- + O_2$

Both the steps are bimolecular elementary reactions. Species IO^- is called as an intermediate since it is formed during the course of the reaction but not in the overall balanced equation. The first step, being slow, is the rate determining step. Thus, the rate of formation of intermediate will determine the rate of this reaction.

Thus, from the discussion, till now, we conclude the following:

(i) Order of a reaction is an experimental quantity. It can be zero and even a fraction but molecularity cannot be zero or a non integer.

(ii) Order is applicable to elementary as well as complex reactions whereas molecularity is applicable only for elementary reactions. For complex reaction molecularity has no meaning.

(iii) For complex reaction, order is given by the slowest step and generally, molecularity of the slowest step is same as the order of the overall reaction.

3.10 Temperature Dependence of the Rate of a Reaction

Most of the chemical reactions are accelerated by increase in temperature. For example, in decomposition of N_2O_5, the time taken for half of the original amount of material to decompose is 12 min at 50°C, 5 h at 25°C and 10 days at 0°C. You also know that in a mixture of potassium permanganate ($KMnO_4$) and oxalic acid ($H_2C_2O_4$), potassium permanganate gets decolourised faster at a higher temperature than that at a lower temperature.

It has been found that **for a chemical reaction with rise in temperature by 10°, the rate constant is nearly doubled.**

The temperature dependence of the rate of a chemical reaction can be accurately explained by Arrhenius equation. It was first proposed by Dutch chemist, J.H. van't Hoff but Swedish chemist, Arrhenius provided its physical justification and interpretation.

$$k = A\, e^{-Ea/RT}$$

where A is the Arrhenius factor or the frequency factor. It is also called pre-exponential factor. It is a constant specific to a particular reaction.

R is gas constant and E_a is activation energy measured in joules/mole (J mol^{-1}).

It can be understood clearly using the following simple reaction

$$H_2(g) + I_2(g) \rightarrow 2HI(g)$$

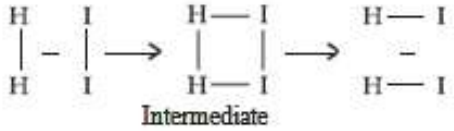


Fig.3.15 Formation of HI through the intermediate

According to Arrhenius, this reaction can take place only when a molecule of hydrogen and a molecule of iodine collide to form an unstable intermediate. It exists for a very short time and then breaks up to form two molecules of hydrogen iodide.

The energy required to form this intermediate, called **activated complex** (C), is known as **activation energy (*E*a)**. Below Fig. 3.16 is obtained by plotting potential energy vs reaction coordinate. Reaction coordinate represents the profile of energy change when reactants change into products.

Some energy is released when the complex decomposes to form products. So, the final heat of the reaction depends upon the nature of reactants and products.

All the molecules in the reacting species do not have the same kinetic energy. Since it is difficult to predict the behaviour of any one molecule with precision, Ludwig Boltzmann and James Clark Maxwell used statistics to predict the behaviour of large number of molecules. According to them, the distribution of kinetic energy may be described by plotting the fraction of molecules (NE/NT) with a given kinetic energy (E) vs kinetic energy (Fig.3.17). Here, NE is the number of molecules with energy E and NT is total number of molecules.

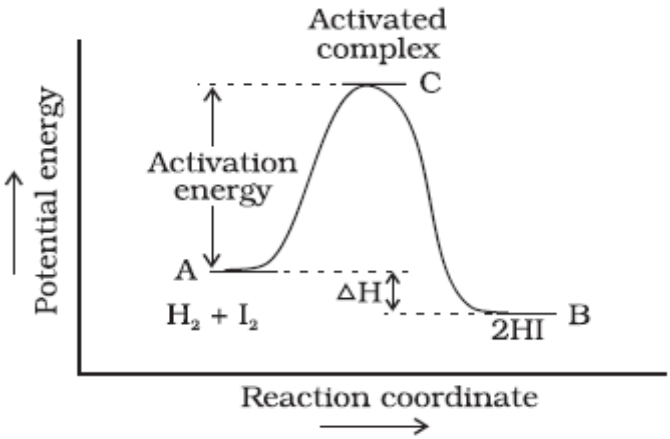


Fig. 3.16 Diagram showing plot of potential energy vs reaction coordinate.

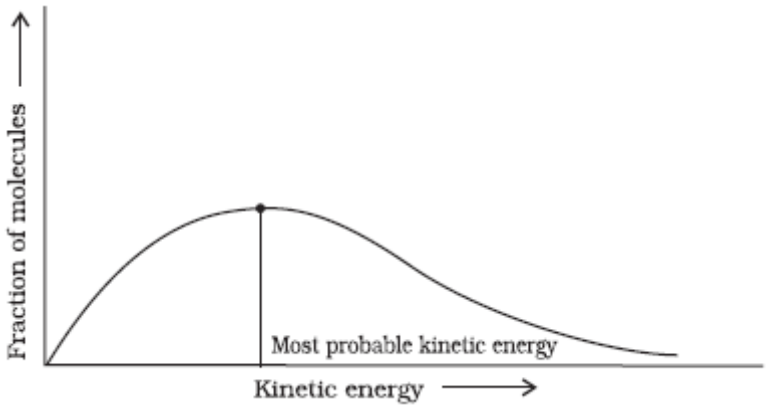


Fig.3.17 Distribution curve showing energies among gaseous molecules

The peak of the curve corresponds to the **most probable kinetic energy,** i.e., kinetic energy of maximum fraction of molecules. There is decreasing number of molecules with energies higher or lower than this value. When the temperature is raised, the maximum of the curve moves to the higher energy value and the curve broadens out, i.e., spreads to the right such that there is a greater proportion of a molecule with much higher energies. The area under the curve must be constant since total probability must be one at all times. We can mark the position of *E*a on Maxwell Boltzmann distribution curve.

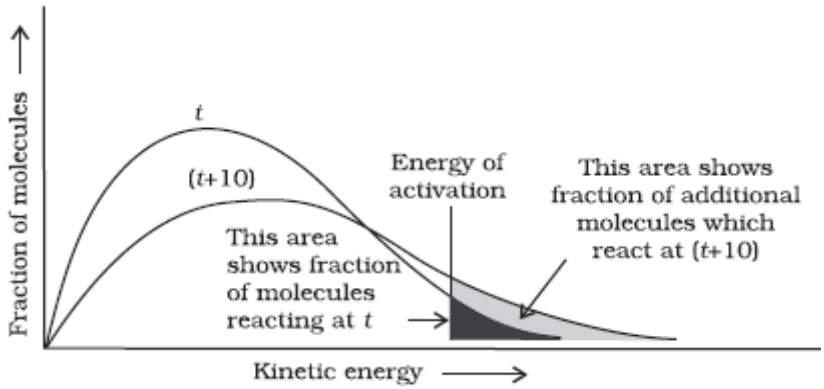


Fig.3.18 Distribution curve showing temperature dependence of rate of a reaction

Increasing the temperature of the substance increases the fraction of molecules, which collide with energies greater than E_a. It is clear from the diagram that in the curve at (t + 10), the area showing the fraction of molecules having energy equal to or greater than activation energy gets doubled leading to doubling the rate of a reaction.

In the Arrhenius equation the factor $e^{-Ea/RT}$ corresponds to the fraction of molecules that have kinetic energy greater than E_a. Taking natural logarithm of both sides of equation.

$$\ln k = -\frac{E_a}{RT} + \ln A$$

The plot of ln k vs 1/T gives a straight line according to the equation as shown in above figure 3.18.

Thus, it has been found from Arrhenius equation mentioned above that increasing the temperature or decreasing the activation energy will result in an increase in the rate of the reaction and an exponential increase in the rate constant.

In Fig. 3.19 slope = $-\frac{E_a}{R}$ and intercept = ln A. So we can calculate E_a and A using these values.

At temperature T_1, equation (4.19) is

$$\ln k_1 = -\frac{E_a}{RT_1} + \ln A$$

At temperature T_2, equation is

$$\ln k_2 = -\frac{E_a}{RT_2} + \ln A$$

(since A is constant for a given reaction)

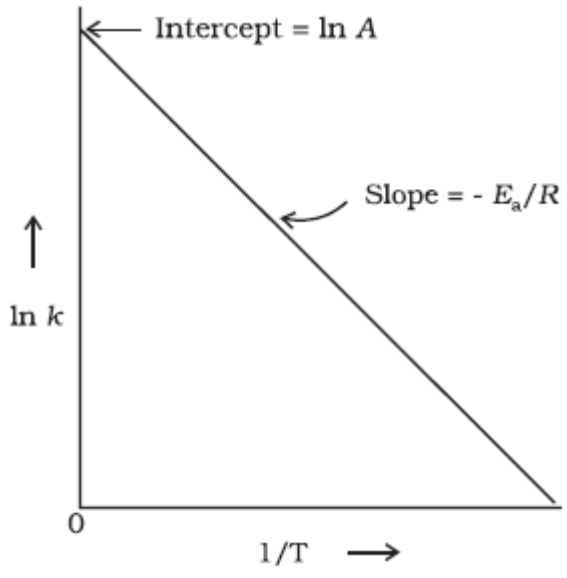


Fig.3.19 A plot between ln k and 1/T

k_1 and k_2 are the values of rate constants at temperatures T_1 and T_2 respectively.

Subtracting equation , we obtain

$$\ln k_2 - \ln k_1 = \frac{E_a}{RT_1} - \frac{E_a}{RT_2}$$

$$\ln \frac{k_2}{k_1} = \frac{E_a}{R}\left[\frac{1}{T_1} - \frac{1}{T_2}\right]$$

$$\log \frac{k_2}{k_1} = \frac{E_a}{2.303R}\left[\frac{1}{T_1} - \frac{1}{T_2}\right]$$

$$\log \frac{k_2}{k_1} = \frac{E_a}{2.303R}\left[\frac{T_2 - T_1}{T_1 T_2}\right]$$

Effect of Catalyst

A catalyst is a substance which alters the rate of a reaction without itself undergoing any permanent chemical change. For example, MnO_2 catalyses, the following reaction so as to increase its rate considerably.

$$2KClO_3 \xrightarrow{MnO_2} 2\ KCl + 3O_2$$

The action of the catalyst can be explained by intermediate complex theory. According to this theory, a catalyst participates in a chemical reaction by forming temporary bonds with the reactants resulting in an intermediate complex. This has a transitory existence and decomposes to yield products and the catalyst.

It is believed that the catalyst provides an alternate pathway or reaction mechanism by reducing the activation energy between reactants and products and hence lowering the potential energy barrier as shown in below Figure 3.20.

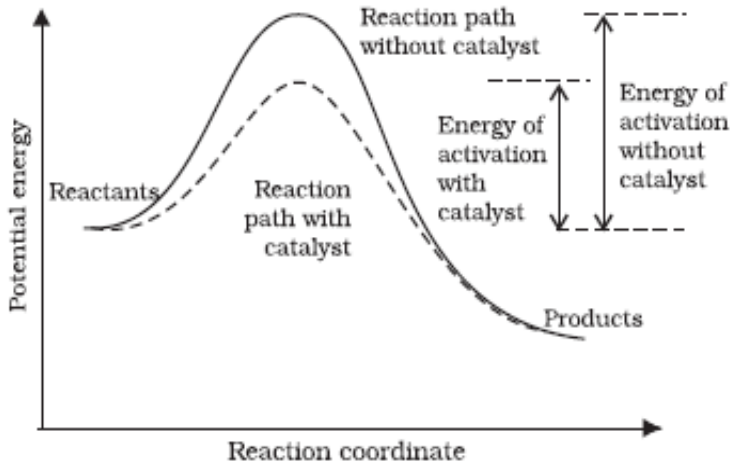


Fig.3.20: Effect of catalyst on activation energy

It is clear from Arrhenius equation that lower the value of activation energy faster will be the rate of a reaction.

A small amount of the catalyst can catalyse a large amount of reactants. A catalyst does not alter Gibbs energy, ΔG of a reaction. It catalyses the spontaneous reactions but does not catalyse non-spontaneous reactions. It is also found that a catalyst does not change the equilibrium constant of a reaction rather, it helps in attaining the equilibrium faster, that is, it catalyses the forward as well as the backward reactions to the same extent so that the equilibrium state remains same but is reached earlier.

3.11 Collision Theory of Chemical Reactions

Though Arrhenius equation is applicable under a wide range of circumstances, collision theory, which was developed by Max Trautz and William Lewis in 1916 -18, provides a greater insight into the energetic and mechanistic aspects of reactions. It is based on kinetic theory of gases. According to this theory, the reactant molecules are assumed to be hard spheres and reaction is postulated to occur when molecules collide with each other. **The number of collisions per second per unit volume of the reaction mixture is known as collision frequency (Z)**. Another factor which affects the rate of chemical reactions is activation energy (as we have already studied). For a bimolecular elementary reaction

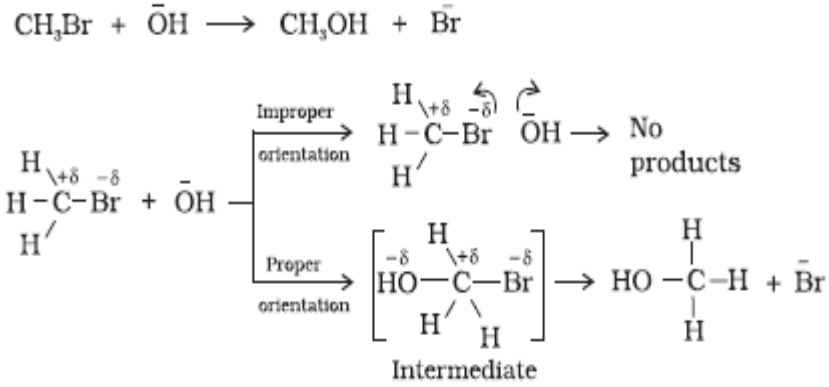
$$A + B \rightarrow \text{Products}$$

rate of reaction can be expressed as

$$\text{Rate} = Z_{AB} e^{-E_a / RT}$$

Where *Z*AB represents the collision frequency of reactants, A and B and e-*E*a /*RT* represents the fraction of molecules with energies equal to or greater than *E*a. Comparing with Arrhenius equation, we can say that *A* is related to collision frequency.

Above Equation predicts the value of rate constants fairly accurately for the reactions that involve atomic species or simple molecules but for complex molecules significant deviations are observed. The reason could be that all collisions do not lead to the formation of products. The collisions in which molecules collide with sufficient kinetic energy (called threshold energy*) and proper orientation, so as to facilitate breaking of bonds between reacting species and formation of new bonds to form products are called as **effective collisions.**



For example, formation of methanol from bromoethane depends upon the orientation of reactant molecules. *The proper orientation of reactant molecules lead to bond formation whereas improper orientation makes them simply bounce back and no products are formed.*

To account for effective collisions, another factor *P*, called the probability or steric factor is introduced. It takes into account the fact that in a collision, molecules must be properly oriented i.e.,

$$\text{Rate} = PZ_{AB} e^{-E_a / RT}$$

Thus, in collision theory activation energy and proper orientation of the molecules together determine the criteria for an effective collision and hence the rate of a chemical reaction.

Collision theory also has certain drawbacks as it considers atoms/ molecules to be hard spheres and ignores their structural aspect. You will study details about this theory and more on other theories in your higher classes.

Summary

Chemical kinetics is the study of chemical reactions with respect to reaction rates, effect of various variables, rearrangement of atoms and formation of intermediates. The rate of a reaction is concerned with decrease in concentration of reactants or increase in the concentration of products per unit time. It can be expressed as instantaneous rate at a particular instant of time and average rate over a large interval of time. A number of factors such as temperature, concentration of reactants, catalyst, affect the rate of a reaction. Mathematical representation of rate of a reaction is given by **rate law**. It has to be determined experimentally and cannot be predicted. **Order of a reaction** with respect to a reactant is the power of its concentration which appears in the rate law equation. The order of a reaction is the sum of all such powers of concentration of terms for different reactants. **Rate constant** is the proportionality factor in the rate law. Rate constant and order of a reaction can be determined from rate law or its integrated rate equation. **Molecularity** is defined only for an elementary reaction. Its values are limited from 1 to 3 whereas order can be 0, 1, 2, 3 or even a fraction. Molecularity and order of an elementary reaction are same.

Temperature dependence of rate constants is described by Arrhenius equation ($k = A\text{e}–E\text{a}/RT$). Ea corresponds to the **activation energy** and is given by the energy difference between activated complex and the reactant molecules, and A (Arrhenius factor or pre-exponential factor) corresponds to the collision frequency. The equation clearly shows that increase of temperature or lowering of Ea will lead to an increase in the rate of reaction and presence of a catalyst lowers the activation energy by providing an alternate path for the reaction. According to collision theory, another factor P called steric factor which refers to the orientation of molecules which collide, is important and contributes to effective collisions, thus, modifying the Arrhenius equation to a / AB $k = P$ Z e$-E\ RT$.

IMPORTANT QUESTIONS

1. Describe the salient features of the Collision Theory of raction rates of Bimolecular reactions.
2. Define order of a reaction, Molecularity of a reaction with suitable examples.
3. What are complex & Disproportionate reactions

CHAPTER 4

SURFACE CHEMISTRY

Surface chemistry deals with phenomena that occur at the surfaces or interfaces. The interface or surface is represented by separating the bulk phases by a hyphen or a slash. For example, the interface between a solid and a gas may be represented by solid-gas or solid/gas. Due to complete miscibility, there is no interface between the gases. The bulk phases that we come across in surface chemistry may be pure compounds or solutions. The interface is normally a few molecules thick but its area depends on the size of the particles of bulk phases. Many important phenomena, noticeable amongst these being corrosion, electrode processes, heterogeneous catalysis, dissolution and crystallisation occur at interfaces. The subject of surface chemistry finds many applications in industry, analytical work and daily life situations.

To accomplish surface studies meticulously, it becomes imperative to have a really clean surface. Under very high vacuum of the order of 10^{-8} to 10^{-9} pascal, it is now possible to obtain ultra clean surface of the metals. Solid materials with such clean surfaces need to be stored in vacuum otherwise these will be covered by molecules of the major components of air namely dioxygen and dinitrogen.

In this Unit, you will be studying some important features of surface chemistry such as adsorption, catalysis and colloids including emulsions and gels.

4.1 Adsorption

There are several examples, which reveal that the surface of a solid has the tendency to attract and retain the molecules of the phase with which it comes into contact. These molecules remain only at the surface and do not go deeper into the bulk. **The accumulation of molecular species at the surface rather than in the bulk of a solid or liquid is termed adsorption**. The molecular species or substance, which concentrates or accumulates at the surface is termed **adsorbate** and the material on the surface of which the adsorption takes place is called **adsorbent**.

Adsorption is essentially a surface phenomenon. Solids, particularly in finely divided state, have large surface area and therefore, charcoal, silica gel, alumina gel, clay, colloids, metals in finely divided state, etc. act as good adsorbents.

Adsorption in action

(i) If a gas like O_2, H_2, CO, Cl_2, NH_3 or SO_2 is taken in a closed vessel containing powdered charcoal, it is observed that the pressure of the gas in the enclosed vessel decreases. The gas molecules concentrate at the surface of the charcoal, i.e., gases are adsorbed at the surface.

(ii) In a solution of an organic dye, say methylene blue, when animal charcoal is added and the solution is well shaken, it is observed that the filtrate turns colourless. The molecules of the dye, thus, accumulate on the surface of charcoal, i.e., are adsorbed.

(iii) Aqueous solution of raw sugar, when passed over beds of animal charcoal, becomes colourless as the colouring substances are adsorbed by the charcoal.

(iv) The air becomes dry in the presence of silica gel because the water molecules get adsorbed on the surface of the gel. It is clear from the above examples that solid surfaces can hold the gas or liquid molecules by virtue of adsorption. The process of removing an adsorbed substance from a surface on which it is adsorbed is called **desorption**.

Distinction between Adsorption and Absorption

In adsorption, the substance is concentrated only at the surface and does not penetrate through the surface to the bulk of the adsorbent, while in absorption, the substance is uniformly distributed throughout the bulk of the solid. For example, when a chalk stick is dipped in ink, the surface retains the colour of the ink due to adsorption of coloured molecules while the solvent of the ink goes deeper into the stick due to absorption. On breaking the chalk stick, it is found to be white from inside. A distinction can be made between absorption and adsorption by taking an example of water vapour. Water vapours are absorbed by anhydrous calcium chloride but adsorbed by silica gel. In other words, in adsorption the concentration of the adsorbate increases only at the surface of the adsorbent, while in absorption the concentration is uniform throughout the bulk of the solid.

Both adsorption and absorption can take place simultaneouslyalso. The term sorption is used to describe both the processes.

Mechanism of Adsorption

Adsorption arises due to the fact that the surface particles of the adsorbent are not in the same environment as the particles inside the bulk. Inside the adsorbent all the forces acting between the particles are mutually balanced but on the surface the particles are not surrounded by atoms or molecules of their kind on all sides, and hence they possess unbalanced or residual attractive forces. These forces of the adsorbent are responsible for attracting the adsorbate particles on its surface.The extent of adsorption increases with the increase of surface area per unit mass of the adsorbent at a given temperature and pressure.

Another important factor featuring adsorption is the heat of adsorption. During adsorption, there is always a decrease in residual forces of the surface, i.e., there is decrease in surface energy which appears as heat. Adsorption, therefore, is invariably an exothermic process. In other words, ΔH of adsorption is always negative. When a gas is adsorbed, the freedom of movement of its molecules becomes restricted. This amount to decrease in the entropy of the gas after adsorption, i.e., ΔS is negative. Adsorption is thus accompanied by decrease in enthalpy as well as decrease in entropy of the system. For a process to be spontaneous, the thermodynamic requirement is that, at constant temperature and pressure, ΔG must be negative, i.e., there is a decrease in Gibbs energy. On the basis of equation, $\Delta G = \Delta H - T\Delta S$, ΔG can be negative if ΔH has sufficiently high negative value as $- T\Delta S$ is positive. Thus, in an adsorption process, which is spontaneous, a combination of these two factors makes ΔG negative. As the adsorption proceeds, ΔH becomes less and less negative ultimately ΔH becomes equal to $T\Delta S$ and ΔG becomes zero. At this state equilibrium is attained.

Types of Adsorption

There are mainly two types of adsorption of gases on solids. If accumulation of gas on the surface of a solid occurs on account of weak van der Waals' forces, the adsorption is termed as **physical adsorption or physisorption**. When the gas molecules or atoms are held to the solid surface by chemical bonds, the adsorption is termed **chemical adsorption or chemisorption**. The chemical bonds may be covalent or ionic in nature. Chemisorption involves a high energy of activation and is, therefore, often referred to as activated adsorption. Sometimes these two processes occur simultaneously and it is not easy to ascertain the type of adsorption. A physical adsorption at low temperature may pass into chemisorption as the temperature is increased. For example, dihydrogen is first adsorbed on nickel by van der Waals' forces. Molecules of hydrogen then dissociate to form hydrogen atoms which are held on the surface by chemisorption.

Some of the important characteristics of both types of adsorption are described below and Comparison of Physisorption and Chemisorption is given Table. 4.1.

Characteristics of physisorption

(i) *Lack of specificity:* A given surface of an adsorbent does not show any preference for a particular gas as the van der Waals' forces are universal.

(ii) *Nature of adsorbate:* The amount of gas adsorbed by a solid depends on the nature of gas. In general, easily liquefiable gases (i.e., with higher critical temperatures) are readily adsorbed as van der Waals' forces are stronger near the critical temperatures. Thus, 1g of activated charcoal adsorbs more sulphur dioxide (critical temperature 630K), than methane (critical temperature 190K) which is still more than 4.5 mL of dihydrogen (critical temperature 33K).

(iii) *Reversible nature:* Physical adsorption of a gas by a solid is generally reversible. Thus, Solid + Gas l Gas/Solid + Heat More of gas is adsorbed when pressure is increased as the volume of the gas decreases (Le–Chateliers's principle) and the gas can be removed by decreasing pressure. Since the adsorption process is exothermic, the physical adsorption occurs readily at low temperature and decreases with increasing temperature (Le-Chatelier's principle).

(iv) *Surface area of adsorbent:* The extent of adsorption increases with the increase of surface area of the adsorbent. Thus, finely divided metals and porous substances having large surface areas are good adsorbents.

(v) *Enthalpy of adsorption:* No doubt, physical adsorption is an exothermic process but its enthalpy of adsorption is quite low (20–40 kJ mol^{-1}). This is because the attraction between gas molecules and solid surface is only due to weak van der Waals' forces.

Characteristics of chemisorption

(i) *High specificity:* Chemisorption is highly specific and it will only occur if there is some possibility of chemical bonding between adsorbent and adsorbate. For example, oxygen is adsorbed on metals by virtue of oxide formation and hydrogen is adsorbed by transition metals due to hydride formation.

(ii) *Irreversibility:* As chemisorption involves compound formation, it is usually irreversible in nature. Chemisorption is also an exothermic process but the process is very slow at low temperatures on account of high energy of activation. Like most chemical changes, adsorption often increases with rise of temperature. Physisorption of a gas adsorbed at low temperature may change into chemisorption at a high temperature. Usually high pressure is also favourable for chemisorption.

(iii) *Surface area:* Like physical adsorption, chemisorption also increases with increase of surface area of the adsorbent.

(iv) *Enthalpy of adsorption:* Enthalpy of chemisorption is high (80-240 kJ mol^{-1}) as it involves chemical bond formation.

Table 4.1 Comparison of Physisorption and Chemisorption

Physisorption	Chemisorption
1. It arises because of van der Waals' forces.	1. It is caused by chemical bond formation.
2. It is not specific in nature.	2. It is highly specific in nature.
3. It is reversible in nature.	3. It is irreversible.
4. It depends on the nature of gas. More easily liquefiable gases are adsorbed readily.	4. It also depends on the nature of gas. Gases which can react with the adsorbent show chemisorptions.
5. Enthalpy of adsorption is low (20-40 kJ mol^{-1}) in this case.	5. Enthalpy of adsorption is high (80-240 kJ mol^{-1}) in this case.
6. Low temperature is favourable for adsorption. It decreases with increase of temperature.	6. High temperature is favourable for adsorption. It increases with the increase of temperature.
7. No appreciable activation energy is needed.	7. High activation energy is sometimes needed.
8. It depends on the surface area. It increases with an increase of surface area.	8. It also depends on the surface area. It too increases with an increase of surface area.
9. It results into multimolecular layers on adsorbent surface under high pressure.	9. It results into unimolecular layer.

Adsorption Isotherms

The variation in the amount of gas adsorbed by the adsorbent with pressure at constant temperature can be expressed by means of a curve termed as **adsorption isotherm**.

Freundlich adsorption isotherm: Freundlich, in 1909, gave an empirical relationship between the quantity of gas adsorbed by unit mass of solid adsorbent and pressure at a particular temperature. The relationship can be expressed by the following equation:

$$\frac{x}{m} = k.P^{1/n} \; (n > 1)$$

where x is the mass of the gas adsorbed on mass m of the adsorbent at pressure P, k and n are constants which depend on the nature of the adsorbent and the gas at a particular temperature. The relationship is generally represented in the form of a curve where mass of the gas adsorbed per gram of the adsorbent is plotted against pressure. These curves indicate that at a fixed pressure, there is a decrease in physical adsorption with increase in temperature. These curves always seem to approach saturation at high pressure.

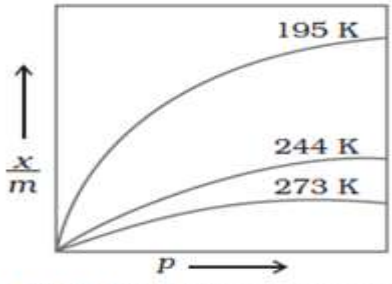


Fig. 4.1 Adorption isotherm

Taking logarithm of eq.

$$\log \frac{x}{m} = \log k + \frac{1}{n} \log P$$

The validity of Freundlich isotherm can be verified by plotting log *xm*on *y*-axis (ordinate) and log *x/m P* on *x*-axis (abscissa). If it comes to be a straight line, the Freundlich isotherm is valid, otherwise not. The slope of the straight line gives the value of 1/*n*. The intercept on the *y*-axis gives the value of log *k*.

Freundlich isotherm explains the behavior of adsorption in an approximate manner. The factor 1/*n* can have values between 0 and 1 (probable range 0.1 to 0.5). Thus, equation holds good over a limited range of pressure.

When 1/*n* = 0x *x/m* = constant, the adsorption is independent of pressure.

When $\frac{1}{n} = 1, \frac{x}{m} = k\,P$, i.e. $\frac{x}{m} \propto P$, the adsorption varies directly with pressure.

Both the conditions are supported by experimental results. The experimental isotherms always seem to approach saturation at high pressure. This cannot be explained by Freundlich isotherm. Thus, it fails at high pressure.

Adsorption from Solution Phase

Solids can adsorb solutes from solutions also. When a solution of acetic acid in water is shaken with charcoal, a part of the acid is adsorbed by the charcoal and the concentration of the acid decreases in the solution. Similarly, the litmus solution when shaken with charcoal becomes colourless. The precipitate of Mg(OH)2 attains blue colour when precipitated in presence of magneson reagent. The colour is due to adsorption of magneson. The following observations have been made in the case of adsorption from solution phase:

(i) The extent of adsorption decreases with an increase in temperature.
(ii) The extent of adsorption increases with an increase of surface area of the adsorbent.
(iii) The extent of adsorption depends on the concentration of the solute in solution.
(iv) The extent of adsorption depends on the nature of the adsorbent and the adsorbate.

The precise mechanism of adsorption from solution is not known. Freundlich's equation approximately describes the behaviour of adsorption from solution with a difference that instead of pressure, concentration of the solution is taken into account, i.e.,

$$\frac{x}{m} = kC^{1/n}$$

(*C* is the equilibrium concentration, i.e., when adsorption is complete). On taking logarithm of the above equation, we have

$$\log \frac{x}{m} = \log k + \frac{1}{n} \log C$$

Plotting log *x/m* against log *C* a straight line is obtained which shows the validity of Freundlich isotherm. This can be tested experimentally by taking solutions of different concentrations of acetic acid. Equal volumes of solutions are added to equal amounts of charcoal in different flasks. The final concentration is determined in each flask after adsorption. The difference in the initial and final concentrations give the value of *x*. Using the above equation, validity of Freundlich isotherm can be established.

Applications of Adsorption

The phenomenon of adsorption finds a number of applications. Important ones are listed here:

(i) *Production of high vacuum*: The remaining traces of air can be adsorbed by charcoal from a vessel evacuated by a vacuum pump to give a very high vacuum.

(ii) *Gas masks*: Gas mask (a device which consists of activated charcoal or mixture of adsorbents) is usually used for breathing in coal mines to adsorb poisonous gases.

(iii) *Control of humidity*: Silica and aluminium gels are used as adsorbents for removing moisture and controlling humidity.

(iv) *Removal of colouring matter from solutions*: Animal charcoal removes colours of solutions by adsorbing coloured impurities.

(v) *Heterogeneous catalysis*: Adsorption of reactants on the solid surface of the catalysts increases the rate of reaction. There are many gaseous reactions of industrial importance involving solid catalysts. Manufacture of ammonia using iron as a catalyst, manufacture of H_2SO_4 by contact process and use of finely divided nickel in the hydrogenation of oils are excellent examples of heterogeneous catalysis.

(vi) *Separation of inert gases*: Due to the difference in degree of adsorption of gases by charcoal, a mixture of noble gases can be separated by adsorption on coconut charcoal at different temperatures.

(vii) *In curing diseases*: A number of drugs are used to kill germs by getting adsorbed on them.

(viii) *Froth floatation process*: A low grade sulphide ore is concentrated by separating it from silica and other earthy matter by this method using pine oil and frothing agent (see Unit 6).

(ix) *Adsorption indicators*: Surfaces of certain precipitates such as silver halides have the property of adsorbing some dyes like eosin, fluorescein, etc. and thereby producing a characteristic colour at the end point.

(x) *Chromatographic analysis*: Chromatographic analysis based on the phenomenon of adsorption finds a number of applications in analytical and industrial fields.

4.2 Catalysis

Potassium chlorate, when heated strongly decomposes slowly giving dioxygen. The decomposition occurs in the temperature range of 653-873K.

$$2KClO_3 \rightarrow 2KCl + 3O_2$$

However, when a little of manganese dioxide is added, the decomposition takes place at a considerably lower temperature range, i.e., 473-633K and also at a much accelerated rate. The added manganese dioxide remains unchanged with respect to its mass and composition. In a similar manner, the rates of a number of chemical reactions can be altered by the mere presence of a foreign substance.

The systematic study of the effect of various foreign substances on the rates of chemical reactions was first made by Berzelius, in 1835. He suggested the term **catalyst** for such substances. Substances, which alter the rate of a chemical reaction and themselves remain chemically and quantitatively unchanged after the reaction, are known as catalysts, and the phenomenon is known as catalysis. You have already studied about catalysts and its functioning in Section 4.5.

Promoters and poisons

Promoters are substances that enhance the activity of a catalyst while poisons decrease the activity of a catalyst. For example, in Haber's process for manufacture of ammonia, molybdenum acts as a promoter for iron which is used as a catalyst.

$$N_2(g) + 3H_2(g) \xrightarrow[Mo(s)]{Fe(s)} 2NH_3(g)$$

Auto Catalysis

During a chemical reaction if one of the products formed acts as a catalyst, the phenomenon is called autocatalysis. Some of the examples are:

During the titration of oxalic acid with $KMnO_4$ solution in the presence of dil. H_2SO_4, the color of $KMnO_4$ fades slowly at the start but fades fast later due to the formation of Mn^{2+}, which acts as autocatalyst.

$$2KMnO_4 + 3\,H_2SO_4 + 5\,H_2C_2O_4 \longrightarrow K_2SO_4 + 2MnSO_4 + 8H_2O + 10CO_2$$

Catalysis can be broadly divided into two groups

4.3 Homogeneous and Heterogeneous Catalysis

Catalysis can be broadly divided into two groups:

(a) Homogeneous catalysis

When the reactants and the catalyst are in the same phase (i.e.,liquid or gas), the process is said to be homogeneous catalysis. The following are some of the examples of homogeneous catalysis:

(i) Oxidation of sulphur dioxide into sulphur trioxide with dioxygen in the presence of oxides of nitrogen as the catalyst in the lead chamber process.

$$2SO_2(g) + O_2(g) \xrightarrow{NO(g)} 2SO_3(g)$$

The reactants, sulphur dioxide and oxygen, and the catalyst, nitric oxide, are all in the same phase.

(ii) Hydrolysis of methyl acetate is catalysed by H+ ions furnished by hydrochloric acid.

$$CH_3COOCH_3(l) + H_2O(l) \xrightarrow{HCl(l)} CH_3COOH(aq) + CH_3OH(aq)$$

Both the reactants and the catalyst are in the same phase.

(iii) Hydrolysis of sugar is catalysed by H+ ions furnished by sulphuric acid.

$$\underset{\text{Solution}}{C_{12}H_{22}O_{11}(aq)} + H_2O(l) \xrightarrow{H_2SO_4(l)} \underbrace{\underset{\text{Glucose}}{C_6H_{12}O_6(aq)} + \underset{\text{Fructose}}{C_6H_{12}O_6(aq)}}_{\text{Solution}}$$

Both the reactants and the catalyst are in the same phase.

(b) Heterogeneous catalysis

The catalytic process in which the reactants and the catalyst are in different phases is known as heterogeneous catalysis. Some of the examples of heterogeneous catalysis are given below:

(i) Oxidation of sulphur dioxide into sulphur trioxide in the presence of Pt.

$$2SO_2(g) \xrightarrow{Pt(s)} 2SO_3(g)$$

The reactant is in gaseous state while the catalyst is in the solid state.

(ii) Combination between dinitrogen and dihydrogen to form ammonia in the presence of finely divided iron in Haber's process.

$$N_2(g) + 3H_2(g) \xrightarrow{Fe(s)} 2NH_3(g)$$

The reactants are in gaseous state while the catalyst is in the solid state.

(iii) Oxidation of ammonia into nitric oxide in the presence of platinum gauze in Ostwald's process.

$$4NH_3(g) + 5O_2(g) \xrightarrow{Pt(s)} 4NO(g) + 6H_2O(g)$$

The reactants are in gaseous state while the catalyst is in the solid state.

(iv) Hydrogenation of vegetable oils in the presence of finely divided nickel as catalyst.

$$\text{Vegetable oils(l)} + H_2(g) \xrightarrow{Ni(s)} \text{Vegetable ghee(s)}$$

One of the reactants is in liquid state and the other in gaseous state while the catalyst is in the solid state.

Adsorption Theory of Heterogeneous Catalysis

This theory explains the mechanism of heterogeneous catalysis. The old theory, known as adsorption theory of catalysis, was that the reactants in gaseous state or in solutions, are adsorbed on the surface of the solid catalyst. The increase in concentration of the reactants on the surface increases the rate of reaction. Adsorption being an exothermic process, the heat of adsorption is utilized in enhancing the rate of the reaction.

The catalytic action can be explained in terms of the intermediate compound formation, the theory of which you have already studied in Section 4.5.1

The modern adsorption theory is the combination of intermediate compound formation theory and the old adsorption theory. The catalytic activity is localized on the surface of the catalyst. The mechanism involves five steps:

(i) Diffusion of reactants to the surface of the catalyst.

(ii) Adsorption of reactant molecules on the surface of the catalyst.

(iii) Occurrence of chemical reaction on the catalyst's surface through formation of an intermediate (Fig.4.1).

(iv) Desorption of reaction products from the catalyst surface, and thereby, making the surface available again for more reaction to occur.

(v) Diffusion of reaction products away from the catalyst's surface. The surface of the catalyst unlike the inner part of the bulk, has free valencies which provide the seat for chemical forces of attraction. When a gas comes in contact with such a surface, its molecules are held up there due to loose chemical combination. If different molecules are adsorbed side by side, they may react with each other resulting in the formation of new molecules. Thus, formed molecules may evaporate leaving the surface for the fresh reactant molecules.

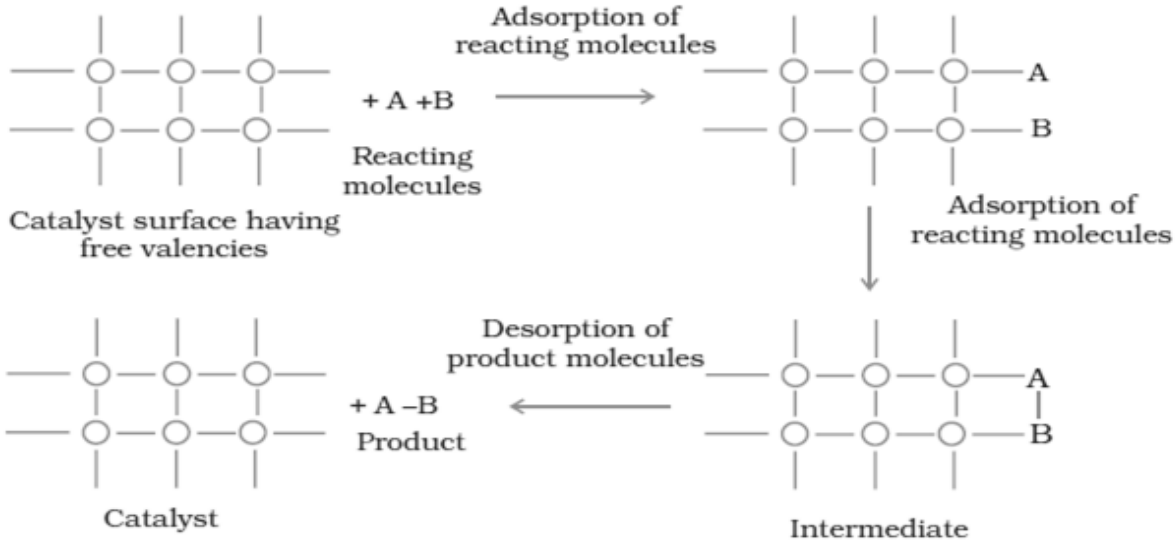


Fig.4.1 Adsorption of reacting molecules, formation of intermediate and desorption of products

This theory explains why the catalyst remains unchanged in mass and chemical composition at the end of the reaction and is effective even in small quantities. It however, does not explain the action of catalytic promoters and catalytic poisons.

Important features of solid catalysts

(a) Activity

The activity of a catalyst depends upon the strength of chemisorptions to a large extent. The reactants must get adsorbed reasonably strongly on to the catalyst to become active. However, they must not get adsorbed so strongly that they are immobilised and other reactants are left with no space on the catalyst's surface for adsorption. It has been found that for hydrogenation reaction, the catalytic activity increases from Group 5 to Group 11 metals with maximum activity being shown by groups 7-9 elements of the periodic table (Class XI, Unit 3).

$$2H_2(g) + O_2(g) \xrightarrow{Pt} 2H_2O(l)$$

(b) Selectivity

The selectivity of a catalyst is its ability to direct a reaction to yield a particular product. For example, starting with H_2 and CO, and using different catalysts, we get different products.

$$\text{(i)}\quad CO(g) + 3H_2(g) \xrightarrow{Ni} CH_4(g) + H_2O(g)$$

$$\text{(ii)}\quad CO(g) + 2H_2(g) \xrightarrow{Cu/ZnO\text{-}Cr_2O_3} CH_3OH(g)$$

$$\text{(iii)}\quad CO(g) + H_2(g) \xrightarrow{Cu} HCHO(g)$$

Thus, it can be inferred that the action of a catalyst is highly selective in nature, i.e., a given substance can act as a catalyst only in a particular reaction and not for all the reactions. It means that a substance which acts as a catalyst in one reaction may fail to catalyse another reaction.

Shape-Selective Catalysis by Zeolites

The catalytic reaction that depends upon the pore structure of the catalyst and the size of the reactant and product molecules is called **shape-selective catalysis.** Zeolites are good shape-selective catalysts because of their honeycomb-like structures. They are microporous aluminosilicates with three dimensional network of silicates in which some silicon atoms are replaced by aluminium atoms giving A_l–O–Si framework. The reactions taking place in zeolites depend upon the size and shape of reactant and product molecules as well as upon the pores and cavities of the zeolites. They are found in nature as well as synthesised for catalytic selectivity.

Zeolites are being very widely used as catalysts in petrochemical industries for cracking of hydrocarbons and isomerisation. An important zeolite catalyst used in the petroleum industry is ZSM-5. It converts alcohols directly into gasoline (petrol) by dehydrating them to give a mixture of hydrocarbons.

Enzyme Catalysis

Enzymes are complex nitrogenous organic compounds which are produced by living plants and animals. They are actually protein molecules of high molecular mass and form colloidal solutions in water. They are very effective catalysts; catalyse numerous reactions, especially those connected with natural processes. Numerous reactions that occur in the bodies of animals and plants to maintain the life process are catalysed by enzymes. The enzymes are, thus, termed as **biochemical catalysts** and the phenomenon is known as **biochemical catalysis**.

Many enzymes have been obtained in pure crystalline state from living cells. However, the first enzyme was synthesised in the laboratory in 1969. The following are some of the examples of enzyme-catalysed reactions:

(i) *Inversion of cane sugar*: The invertase enzyme converts cane sugar into glucose and fructose.

$$\underset{\text{Cane sugar}}{C_{12}H_{22}O_{11}(aq)} + H_2O(l) \xrightarrow{\text{Invertase}} \underset{\text{Glucose}}{C_6H_{12}O_6(aq)} + \underset{\text{Fructose}}{C_6H_{12}O_6(aq)}$$

(ii) *Conversion of glucose into ethyl alcohol*: The zymase enzyme converts glucose into ethyl alcohol and carbon dioxide.

$$\underset{\text{Starch}}{2(C_6H_{10}O_5)_n(aq)} + nH_2O(l) \xrightarrow{\text{Diastase}} \underset{\text{Maltose}}{nC_{12}H_{22}O_{11}(aq)}$$

(iii) *Conversion of starch into maltose*: The diastase enzyme converts starch into maltose.

$$\underset{\text{Starch}}{2(C_6H_{10}O_5)_n(aq)} + nH_2O(l) \xrightarrow{\text{Diastase}} \underset{\text{Maltose}}{nC_{12}H_{22}O_{11}(aq)}$$

(iv) *Conversion of maltose into glucose*: The maltase enzyme converts maltose into glucose.

$$\underset{\text{Maltose}}{C_{12}H_{22}O_{11}(aq)} + H_2O(l) \xrightarrow{\text{Maltase}} \underset{\text{Glucose}}{2C_6H_{12}O_6(aq)}$$

(v) *Decomposition of urea into ammonia and carbon dioxide*: The enzyme urease catalyses this decomposition.

$$NH_2CONH_2(aq) + H_2O(l) \xrightarrow{\text{Urease}} 2NH_3(g) + CO_2(g)$$

(vi) In stomach, the pepsin enzyme converts proteins into peptides while in intestine, the pancreatic trypsin converts proteins into amino acids by hydrolysis.

(vii) *Conversion of milk into curd*: It is an enzymatic reaction brought about by lacto bacilli enzyme present in curd.

Below Table 4.1 gives the summary of some important enzymatic reactions.

Table 4.1 Some Enzymatic Reactions

Enzyme	Source	Enzymatic reaction
Invertase	Yeast	Sucrose → Glucose and fructose
Zymase	Yeast	Glucose → Ethyl alcohol and carbon dioxide
Diastase	Malt	Starch → Maltose
Maltase	Yeast	Maltose → Glucose
Urease	Soyabean	Urea → Ammonia and carbon dioxide
Pepsin	Stomach	Proteins → Amino acids

Characteristics of enzyme catalysis

Enzyme catalysis is unique in its efficiency and high degree of specificity.

The following characteristics are exhibited by enzyme catalysts:

(i) *Most highly efficient*: One molecule of an enzyme may transform one million molecules of the reactant per minute.

(ii) *Highly specific nature*: Each enzyme is specific for a given reaction, i.e., one catalyst cannot catalyse more than one reaction. For example, the enzyme urease catalyses the hydrolysis of urea only. It does not catalyse hydrolysis of any other amide.

(iii) *Highly active under optimum temperature*: The rate of an enzyme reaction becomes maximum at a definite temperature, called the optimum temperature. On either side of the optimum temperature, the enzyme activity decreases. The optimum temperature range for enzymatic activity is 298-310K. Human body temperature being 310 K is suited for enzyme-catalysed reactions.

(iv) *Highly active under optimum pH*: The rate of an enzyme-catalysed reaction is maximum at a particular pH called optimum pH, which is between pH values 5-7.

(v) *Increasing activity in presence of activators and co-enzymes*: The enzymatic activity is increased in the presence of certain substances, known as co-enzymes. It has been observed that when a small non-protein (vitamin) is present along with an enzyme, the catalytic activity is enhanced considerably. Activators are generally metal ions such as Na^+, Mn^{2+}, Co^{2+}, Cu^{2+}, etc. These metal ions, when weakly bonded to enzyme molecules, increase their catalytic activity. Amylase in presence of sodium chloride i.e., Na^+ ions are catalytically very active.

(vi) *Influence of inhibitors and poisons*: Like ordinary catalysts, enzymes are also inhibited or poisoned by the presence of certain substances. The inhibitors or poisons interact with the active functional groups on the enzyme surface and often reduce or completely destroy the catalytic activity of the enzymes. The use of many drugs is related to their action as enzyme inhibitors in the body.

Mechanism of enzyme catalysis

There are a number of cavities present on the surface of colloidal particles of enzymes. These cavities are of characteristic shape and possess active groups such as $-NH_2$, -COOH, -SH, -OH, etc. These are actually the active centres on the surface of enzyme particles. The molecules of the reactant (substrate), which have complementary shape, fit into these cavities just like a key fits into a lock. On account of the presence of active groups, an activated complex is formed which then decomposes to yield the products.

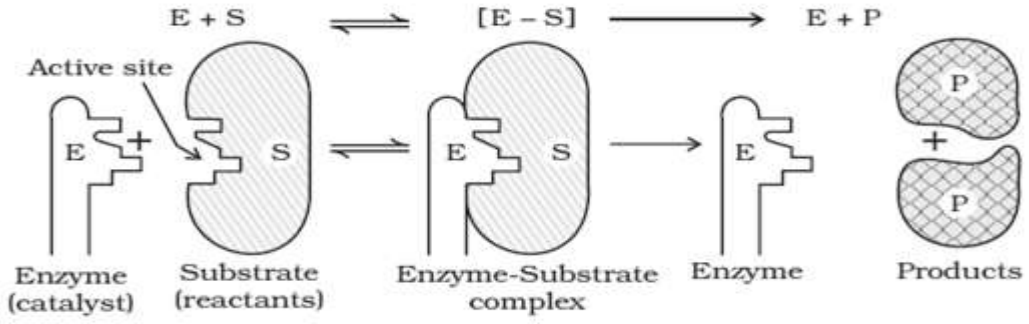


Fig.4.2 Mechanism of enzyme catalysed reaction

Thus, the enzyme-catalysed reactions may be considered to proceed in two steps.

Step 1: Binding of enzyme to substrate to form an activated complex.

$$E + S \rightarrow ES^{\neq}$$

Step 2: Decomposition of the activated complex to form product.

$$ES^{\neq} \rightarrow E + P$$

Catalysts in Industry

Some of the important technical catalytic processes are listed in Table 4.2 to give an idea about the utility of catalysts in industries.

Table 4.2 Some Industrial Catalytic Processes

Process	Catalyst
1. Haber's process for the manufacture of ammonia $N_2(g) + 3H_2(g) \rightarrow 2NH_3(g)$	Finely divided iron, molybdenum as promoter; conditions: 200 bar pressure and 723-773K temperature.
2. Ostwald's process for the manufacture of nitric acid. $4NH_3(g) + 5O_2(g) \rightarrow 4NO(g) + 6H_2O(g)$ $2NO(g) + O_2(g) \rightarrow 2NO_2(g)$ $4NO_2(g) + 2H_2O(l) + O_2(g) \rightarrow 4HNO_3(aq)$	Platinised asbestos; temperature 573K.
3. Contact process for the manufacture of sulphuric acid. $2SO_2(g) + O_2(g) \rightarrow 2SO_3(g)$ $SO_3(g) + H_2SO_4(aq) \rightarrow H_2S_2O_7(l)$ oleum $H_2S_2O_7(l) + H_2O(l) \rightarrow 2H_2SO_4(aq)$	Platinised asbestos or vanadium pentoxide (V_2O_5); temperature 673-723K.

4.3 Colloids

We have learnt in Unit 2 that solutions are homogeneous systems. We also know that sand in water when stirred gives a suspension, which slowly settles down with time. Between the two extremes of suspensions and solutions we come across a large group of systems called colloidal dispersions or simply colloids.

A colloid is a heterogeneous system in which one substance is dispersed (dispersed phase) as very fine particles in another substance called dispersion medium. The essential difference between a solution and a colloid is that of particle size. While in a solution, the constituent particles are ions or small molecules, in a colloid, the dispersed phase may consist of particles of a single macromolecule (such as protein or synthetic polymer) or an aggregate of many atoms, ions or molecules. Colloidal particles are larger than simple molecules but small enough to remain suspended. Their range of diameters is between 1 and 1000 nm (10^{-9} to 10^{-6} m).

Colloidal particles have an enormous surface area per unit mass as a result of their small size. Consider a cube with 1 cm side. It has a total surface area of 6 cm^2. If it were divided equally into 1012 cubes, the cubes would be the size of large colloidal particles and have a total surface area of 60,000 cm^2 or 6 m^2. This enormous surface area leads to some special properties of colloids to be discussed later in this Unit.

Classification of Colloids

Colloids are classified on the basis of the following criteria:

(i) Physical state of dispersed phase and dispersion medium
(ii) Nature of interaction between dispersed phase and dispersion medium
(iii) Type of particles of the dispersed phase.

Classification Based on Physical State of Dispersed Phase and Dispersion Medium

Depending upon whether the dispersed phase and the dispersion medium are solids, liquids or gases, eight types of colloidal systems are possible. A gas mixed with another gas forms a homogeneous mixture and hence is not a colloidal system. The examples of the various types of colloids along with their typical names are listed in below Table 4.3.

Table 4.3 Types of Colloidal Systems

Dispersed phase	Dispersion medium	Type of colloid	Examples
Solid	Solid	Solid sol	Some coloured glasses and gem stones
Solid	Liquid	Sol	Paints, cell fluids
Solid	Gas	Aerosol	Smoke, dust
Liquid	Solid	Gel	Cheese, butter, jellies
Liquid	Liquid	Emulsion	Milk, hair cream
Liquid	Gas	Aerosol	Fog, mist, cloud, insecticide sprays
Gas	Solid	Solid sol	Pumice stone, foam rubber
Gas	Liquid	Foam	Froth, whipped cream, soap lather

Many familiar commercial products and natural objects are colloids. For example, whipped cream is a foam, which is a gas dispersed in a liquid. Firefighting foams, used at emergency airplane landings are also colloidal systems. Most biological fluids are aqueous sols (solids dispersed in water). Within a typical cell, proteins and nucleic acids are colloidal-sized particles dispersed in an aqueous solution of ions and small molecules.

Out of the various types of colloids given in Table 4.3, the most common are **sols** (solids in liquids), **gels** (liquids in solids) and **emulsions** (liquids in liquids). However, in the present Unit, we shall take up discussion of the 'sols' and 'emulsions' only. Further, it may be mentioned that if the dispersion medium is water, the sol is called aquasol or hydrosol and if the dispersion medium is alcohol, it is called alcosol and so on.

Classification Based on Nature of Interaction between Dispersed Phase and Dispersion Medium

Depending upon the nature of interaction between the dispersed phase and the dispersion medium, colloidal sols are divided into two categories, namely, **lyophilic** (solvent attracting) and **lyophobic** (solvent repelling). If water is the dispersion medium, the terms used are hydrophilic and hydrophobic.

(i) *Lyophilic colloids*: The word 'lyophilic' means liquid-loving. Colloidal sols directly formed by mixing substances like gum, gelatine, starch, rubber, etc., with a suitable liquid (the dispersion medium) are called lyophilic sols. An important characteristic of these sols is that if the dispersion medium is separated from the dispersed phase (say by evaporation), the sol can be reconstituted by simply remixing with the dispersion medium. That is why these sols are also called **reversible sols**. Furthermore, these sols are quite stable and cannot be easily coagulated as discussed later.

(ii) *Lyophobic colloids*: The word 'lyophobic' means liquid-hating. Substances like metals, their sulphides, etc., when simply mixed with the dispersion medium do not form the colloidal sol. Their colloidal sols can be prepared only by special methods (as discussed later). Such sols are called lyophobic sols. These sols are readily precipitated (or coagulated) on the addition of small amounts of electrolytes, by heating or by shaking and hence, are not stable. Further, once precipitated, they do not give back the colloidal sol by simple addition of the dispersion medium. Hence, these sols are also called **irreversible sols**. Lyophobic sols need stabilizing agents for their preservation.

4.4 Cleansing action of soaps

It has been mentioned earlier that a micelle consists of a hydrophobic hydrocarbon – like central core. The cleansing action of soap is due to the fact that soap molecules form micelle around the oil droplet in such a way that hydrophobic part of the stearate ions is in the oil droplet and hydrophilic part projects out of the grease droplet like the bristles (Fig.4.3). Since the polar groups can interact with water, the oil droplet surrounded by stearate ions is now pulled in water and removed from the dirty surface. Thus soap helps in emulsification and washing away of oils and fats. The negatively charged sheath around the globules prevents them from coming together and forming aggregates.

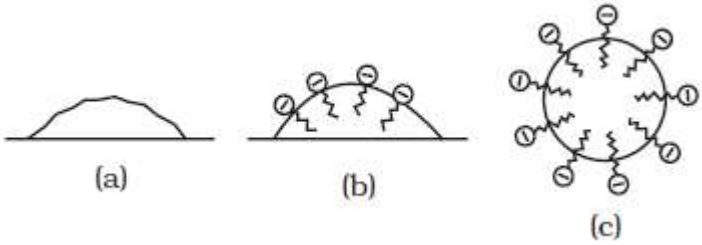


Fig. 4.3 (a) Grease on cloth (b) Stearate ions arranging around the grease droplet and (c) Grease droplet surrounded by stearate ions (micelle formed)

Preparation of Colloids

A few important methods for the preparation of colloids are as follows:

(a) Chemical methods

Colloidal solutions can be prepared by chemical reactions leading to formation of molecules by double decomposition, oxidation, reduction or hydrolysis. These molecules then aggregate leading to formation of sols.

$$As_2O_3 + 3H_2S \xrightarrow{\text{Double decomposition}} As_2S_3(sol) + 3H_2O$$

$$SO_2 + 2H_2S \xrightarrow{\text{Oxidation}} 3S(sol) + 2H_2O$$

$$2\ AuCl_3 + 3\ HCHO + 3H_2O \xrightarrow{\text{Reduction}} 2Au(sol) + 3HCOOH + 6HCl$$

$$FeCl_3 + 3H_2O \xrightarrow{\text{Hydrolysis}} Fe(OH)_3\ (sol) + 3HCl$$

(b) Electrical disintegration or Bredig's Arc method

This process involves dispersion (Fig. 4.4) as well as condensation. Colloidal sols of metals such as gold, silver, platinum, etc., can be prepared

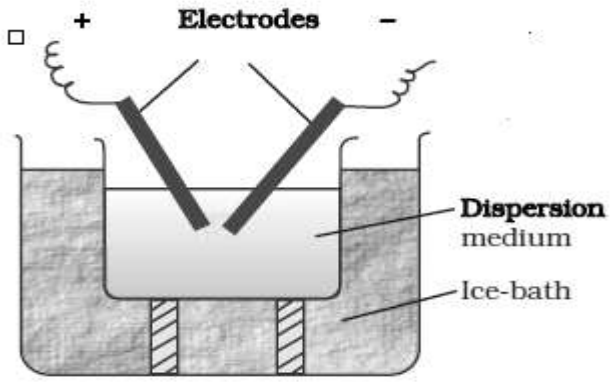


Fig.4.4 Bredig's Arc method

(c) Peptization

Peptization may be defined as the **process of converting a precipitate into colloidal sol** by shaking it with dispersion medium in the presence of a small amount of electrolyte. The electrolyte used for this purpose is called **peptizing agent**. This method is applied, generally, to convert a freshly prepared precipitate into a colloidal sol.

During peptization, the precipitate adsorbs one of the ions of the electrolyte on its surface. This causes the development of positive or negative charge on precipitates, which ultimately break up into smaller particles of the size of a colloid.

Purification of Colloidal Solutions

Colloidal solutions when prepared, generally contain excessive amount of electrolytes and some other soluble impurities. While the presence of traces of electrolyte is essential for the stability of the colloidal solution, larger quantities coagulate it. It is, therefore, necessary to reduce the concentration of these soluble impurities to a requisite minimum. **The process used for reducing the amount of impurities to a requisite minimum is known as purification of colloidal solution**. The purification of colloidal solution is carried out by the following mehods:

(i) *Dialysis*: **It is a process of removing a dissolved substance from a colloidal solution by means of diffusion through a suitable membrane**. Since particles (ions or smaller molecules) in a true solution can pass through animal membrane (bladder) or parchment paper or cellophane sheet but not the colloidal particles, the membrane can be used for dialysis. The apparatus used for this purpose is called **dialyser**. A bag of suitable membrane containing the colloidal solution is suspended in a vessel through which fresh water is continuously flowing. The molecules and ions diffuse through membrane into the outer water and pure colloidal solution is left behind.

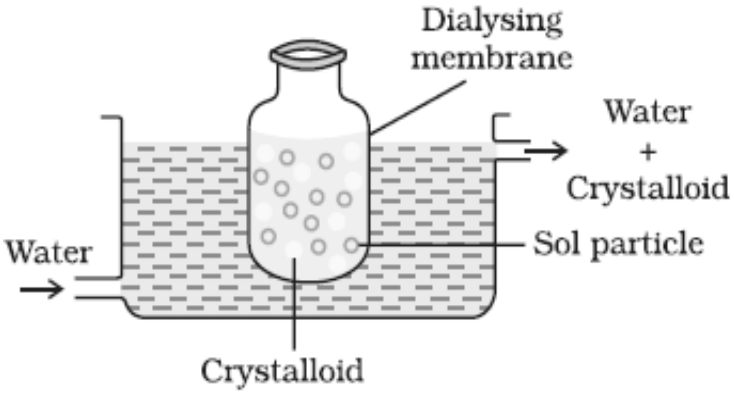


Fig. 4.5 Dialysis

(ii) *Electro-dialysis*: Ordinarily, the process of dialysis is quite slow. It can be made faster by applying an electric field if the dissolved substance in the impure colloidal solution is only an electrolyte. The process is then named electrodialysis. The colloidal solution is placed in a bag of suitable membrane while pure water is taken outside. Electrodes are fitted in the compartment as shown in the below Fig. 4.6. The ions present in the colloidal solution migrate out to the oppositely charged electrodes.

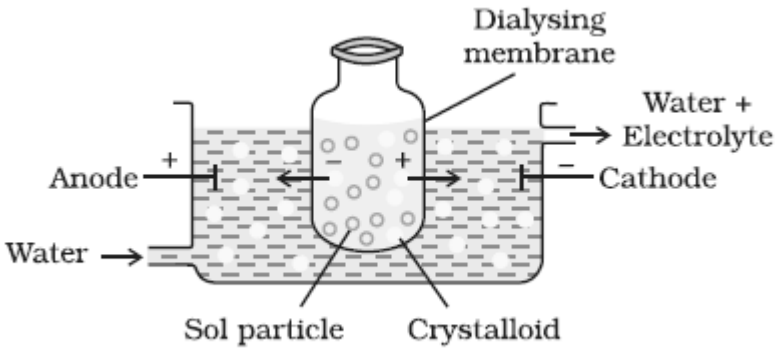


Fig. 4.6. Electro-dialysis

(iii) *Ultrafiltration*: Ultrafiltration is the process of separating the colloidal particles from the solvent and soluble solutes present in the colloidal solution by specially prepared filters, which are permeable to all substances except the colloidal particles. Colloidal particles can pass through ordinary filter paper because the pores are too large. However, the pores of filter paper can be reduced in size by impregnating with **colloidion** solution to stop the flow of colloidal particles. The usual colloidion is a 4% solution of nitrocellulose in a mixture of alcohol and ether. An ultra-filter paper may be prepared by soaking the filter paper in a colloidion solution, hardening by formaldehyde and then finally drying it. Thus, by using ultra-filter paper, the colloidal particles are separated from rest of the materials. Ultrafiltration is a slow process. To speed up the process, pressure or suction is applied. The colloidal particles left on the ultra-filter paper are then stirred with fresh dispersion medium (solvent) to get a pure colloidal solution.

4.5 Properties of Colloidal Solutions

Various properties exhibited by the colloidal solutions are described below:

Tyndall effect

If a homogeneous solution placed in dark is observed in the direction of light, it appears clear and, if it is observed from a direction at right angles to the direction of light beam, it appears perfectly dark. Colloidal solutions viewed in the same way may also appear reasonably clear or translucent by the transmitted light but they show a mild to strong opalescence, when viewed at right angles to the passage of light, i.e., the path of the beam is illuminated by a bluish light. This effect was first observed by Faraday and later studied in detail by Tyndall and is termed as **Tyndall effect (Fig.4.7)**. The bright cone of the light is called **Tyndall cone**. The Tyndall effect is due to the fact that colloidal particles scatter light in all directions in space. This scattering of light illuminates the path of beam in the colloidal dispersion.

Tyndall effect can be observed during the projection of picture in the cinema hall due to scattering of light by dust and smoke particles present there. Tyndall effect is observed only when the following two conditions are satisfied.

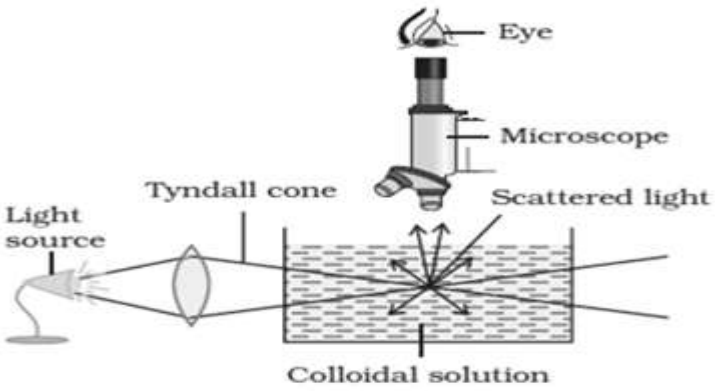


Fig. 4.7 Tyndall effect

(i) The diameter of the dispersed particles is not much smaller than the wavelength of the light used; and

(ii) The refractive indices of the dispersed phase and the dispersion medium differ greatly in magnitude.

Tyndall effect is used to distinguish between a colloidal and true solution. Zsigmondy, in 1903, used Tyndall effect to set up an apparatus known as ultramicroscope. An intense beam of light is focussed on the colloidal solution contained in a glass vessel. The focus of the light is then observed with a microscope at right angles to the beam. Individual colloidal particles appear as bright stars against a dark background. Ultramicroscope does not render the actual colloidal particles visible but only observe the light scattered by them. Thus, ultramicroscope does not provide any information about the size and shape of colloidal particles.

Colour

The colour of colloidal solution depends on the wavelength of light scattered by the dispersed particles. The wavelength of light further depends on the size and nature of the particles. The colour of colloidal solution also changes with the manner in which the observer receives the light. For example, a mixture of milk and water appears blue when viewed by the reflected light and red when viewed by the transmitted light. Finest gold sol is red in colour; as the size of particles increases, it appears purple, then blue and finally golden.

Brownian movement

When colloidal solutions are viewed under a powerful ultramicroscope, the colloidal particles appear to be in a state of continuous zig-zag motion all over the field of view. This motion was first observed by the British botanist, Robert Brown, and is known as Brownian movement (Fig. 4.8). This motion is independent of the nature of the colloid but depends on the size of the particles and viscosity of the solution. Smaller the size and lesser the viscosity, faster is the motion.

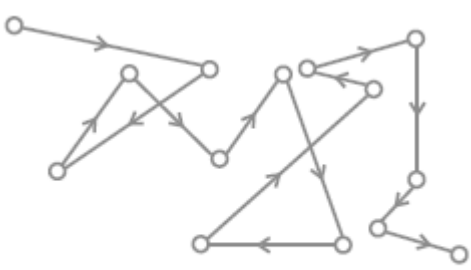


Fig.4.8 Brownian movement

The Brownian movement has been explained to be due to the unbalanced bombardment of the particles by the molecules of the dispersion medium. The Brownian movement has a stirring effect which does not permit the particles to settle and thus, is responsible for the stability of sols.

Charge on colloidal particles

Colloidal particles always carry an electric charge. The nature of this charge is the same on all the particles in a given colloidal solution and may be either positive or negative. A list of some common sols with the nature of charge on their particles is given below:

Positively charged sols	**Negatively charged sols**
Hydrated metallic oxides e.g. $Al_2O_3.xH_2O$, CrO_3 and $Fe_2O_3.xH_2O$ etc.	**Metals e.g. Copper, Silver, gold sols**
Basic dye stuffs e.g. methylene blue sol.	**Metallic sulphides e.g. As_2S_3, Sb_2S_3, CdS sols.**
Haemoglobin (blood)	**Acid dye stuffs, e.g., cosin, congo red sols**
Oxides e.g., TiO_2 sol.	**Sols of starch, gum, gelatine, clay, charcoal etc.**

The charge on the sol particles is due to one or more reasons, viz., due to electron capture by sol particles during electrodispersion of metals, due to preferential adsorption of ions from solution and/or due to formulation of electrical double layer.

Preferential adsorption of ions is the most accepted reason. The sol particles acquire positive or negative charge by preferential adsorption of +ve or –ve ions. When two or more ions are present in the dispersion medium, preferential adsorption of the ion common to the colloidal particle usually takes place. This can be explained by taking the following examples:

(a) When silver nitrate solution is added to potassium iodide solution, the precipitated silver iodide adsorbs iodide ions from the dispersion medium and negatively charged colloidal solution results. However, when KI solution is added to $AgNO_3$ solution, positively charged sol results due to adsorption of Ag+ ions from dispersion medium.

AgI/I^- Negatively charged AgI/Ag^+ Positively charged

(b) If $FeCl_3$ is added to excess of hot water, a positively charged sol of hydrated ferric oxide is formed due to adsorption of Fe^{3+} ions. However, when ferric chloride is added to NaOH a negatively charged sol is obtained with adsorption of OH- ions.

$Fe_2O_3.xH_2O/Fe^{3+}$ Positively charged $Fe_2O_3.xH_2O/OH^-$ Negatively charged

Having acquired a positive or a negative charge by selective adsorption on the surface of a colloidal particle as stated above, this layer attracts counter ions from the medium forming a second layer, as shown below.

$AgI/I^- K^+$ AgI/Ag^+I^-

The combination of the two layers of opposite charges around the colloidal particle is called Helmholtz electrical double layer. According to modern views, the first layer of ions is firmly held and is termed fixed layer while the second layer is mobile which is termed diffused layer. Since separation of charge is a seat of potential, the charges of opposite signs on the fixed and diffused parts of the double layer results in a difference in potential between these layers. This potential difference between the fixed layer and the diffused layer of opposite charges is called the **electrokinetic potential or zeta potential**.

The presence of equal and similar charges on colloidal particles is largely responsible in providing stability to the colloidal solution, because the repulsive forces between charged particles having same charge prevent them from coalescing or aggregating when they come closer to one another.

Electrophoresis: The existence of charge on colloidal particles is confirmed by electrophoresis experiment. When electric potential is applied across two platinum electrodes dipping in a colloidal solution, the colloidal particles move towards one or the other electrode. The movement of colloidal particles under an applied electric potential is called electrophoresis. Positively charged particles move towards the cathode while negatively charged particles move towards the anode. This can be demonstrated by the following experimental setup (Fig. 4.9).

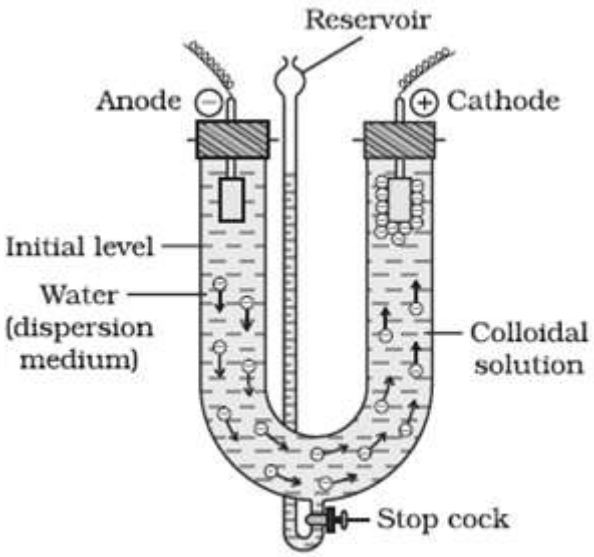


Fig.4.9 Electrophoresis

When electrophoresis, i.e., movement of particles is prevented by some suitable means, it is observed that the dispersion medium begins to move in an electric field. This phenomenon is termed **electroosmosis**.

Coagulation or precipitation: The stability of the lyophobic sols is due to the presence of charge on colloidal particles. If, somehow, the charge is removed, the particles will come nearer to each other to form aggregates (or coagulate) and settle down under the force of gravity.

The process of settling of colloidal particles is called coagulation or precipitation of the sol. The coagulation of the lyophobic sols can be carried out in the following ways:

(i) *By electrophoresis*: The colloidal particles move towards oppositely charged electrodes, get discharged and precipitated.

(ii) *By mixing two oppositely charged sols*: Oppositely charged sols when mixed in almost equal proportions, neutralise their charges and get partially or completely precipitated. Mixing of hydrated ferric oxide (+ve sol) and arsenious sulphide (–ve sol) bring them in the precipitated forms. This type of coagulation is called mutual coagulation.

(iii) *By boiling*: When a sol is boiled, the adsorbed layer is disturbed due to increased collisions with the molecules of dispersion medium. This reduces the charge on the particles and ultimately lead to settling down in the form of a precipitate.

(iv) *By persistent dialysis*: On prolonged dialysis, traces of the electrolyte present in the sol are removed almost completely and the colloids become unstable and ultimately coagulate.

(v) *By addition of electrolytes*: When excess of an electrolyte is added, the colloidal particles are precipitated. The reason is that colloids interact with ions carrying charge opposite to that present on themselves. This causes neutralisation leading to their coagulation. The ion responsible for neutralisation of charge on the particles is called the coagulating ion. A negative ion causes the precipitation of positively charged sol and vice versa. It has been observed that, generally, the greater the valence of the flocculating ion added, the greater is its power to cause precipitation. This is known as Hardy-Schulze rule. In the coagulation of a negative sol, the flocculating power is in the order: $Al^{3+} > Ba^{2+} > Na^{+}$

Similarly, in the coagulation of a positive sol, the flocculating power is in the order:

$$[Fe(CN)_6]^{4-} > PO_4^{3-} > SO_4^{2-} > Cl^-$$

The minimum concentration of an electrolyte in millimoles per litre required to cause precipitation of a sol in two hours is called coagulating value. The smaller the quantity needed, the higher will be the coagulating power of an ion.

Coagulation of lyophilic sols

There are two factors which are responsible for the stability of lyophilic sols. These factors are the charge and solvation of the colloidal particles. When these two factors are removed, a lyophilic sol can be coagulated. This is done (i) by adding an electrolyte and (ii) by adding a suitable solvent. When solvents such as alcohol and acetone are added to hydrophilic sols, the dehydration of dispersed phase occurs. Under this condition, a small quantity of electrolyte can bring about coagulation.

Protection of colloids

Lyophilic sols are more stable than lyophobic sols. This is due to the fact that lyophilic colloids are extensively solvated, i.e., colloidal particles are covered by a sheath of the liquid in which they are dispersed. Lyophilic colloids have a unique property of protecting lyophobic colloids. When a lyophilic sol is added to the lyophobic sol, the lyophilic particles form a layer around lyophobic particles and thus protect the latter from electrolytes. Lyophilic colloids used for this purpose are called protective colloids.

4.6 Emulsions

These are liquid-liquid colloidal systems, i.e., the dispersion of finely divided droplets in another liquid. If a mixture of two immiscible or partially miscible liquids is shaken, a coarse dispersion of one liquid in the other is obtained which is called emulsion. Generally, one of the two liquids is water. There are two types of emulsions. (i) Oil dispersed in water (O/W type) and (ii) Water dispersed in oil (W/O type). In the first system, water acts as

dispersion medium. Examples of this type of emulsion are milk and vanishing cream. In milk, liquid fat is dispersed in water. In the second system, oil acts as dispersion medium. Common examples of this type are butter and cream.

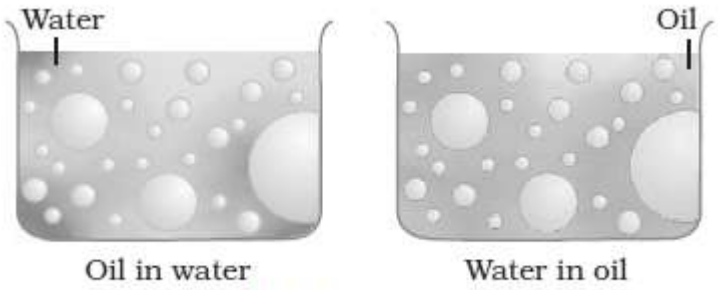


Fig.4.10 Types of emulsions

Emulsions of oil in water are unstable and sometimes they separate into two layers on standing (Fig.4.10). For stabilisation of an emulsion, a third component called emulsifying agent is usually added. The emulsifying agent forms an interfacial film between suspended particles and the medium. The principal emulsifying agents for O/W emulsions are proteins, gums, natural and synthetic soaps, etc., and for W/O, heavy metal salts of fatty acids, long chain alcohols, lampblack, etc.

Emulsions can be diluted with any amount of the dispersion medium. On the other hand, the dispersed liquid when mixed, forms a separate layer. The droplets in emulsions are often negatively charged and can be precipitated by electrolytes. They also show Brownian movement and Tyndall effect. Emulsions can be broken into constituent liquids by heating, freezing, centrifuging, etc.

Applications of colloids

Colloids are widely used in the industry. Following are some examples:

(i) *Electrical precipitation of smoke*: Smoke is a colloidal solution of solid particles such as carbon, arsenic compounds, dust, etc., in air. The smoke, before it comes out from the chimney, is led through a chamber containing plates having a charge opposite to that carried by smoke particles. The particles on coming in contact with these plates lose their charge and get precipitated. The particles thus settle down on the floor of the chamber. The precipitator is called Cottrell precipitator (Fig.4.11).

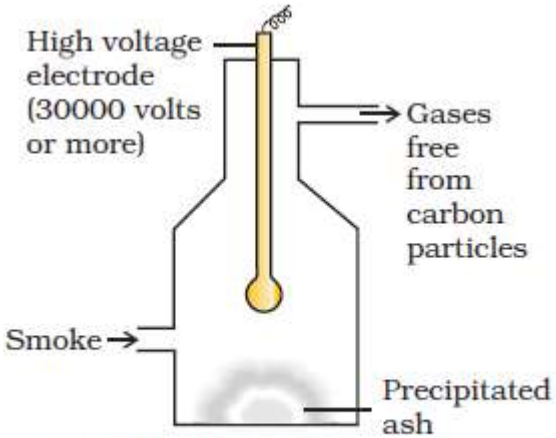


Fig.4.11 Cottell smoke precipitor

(ii) *Purification of drinking water*: The water obtained from natural sources often contains suspended impurities. Alum is added to such water to coagulate the suspended impurities and make water fit for drinking purposes.

(iii) *Medicines*: Most of the medicines are colloidal in nature. For example, argyrol is a silver sol used as an eye lotion. Colloidal antimony is used in curing kalaazar. Colloidal gold is used for intramuscular injection. Milk of magnesia, an emulsion, is used for stomach disorders. Colloidal medicines are more effective because they have

(iv) *Tanning*: Animal hides are colloidal in nature. When a hide, which has positively charged particles, is soaked in tannin, which contains negatively charged colloidal particles, mutual coagulation takes place. This results in the hardening of leather. This process is termed as tanning. Chromium salts are also used in place of tannin.

(v) *Cleansing action of soaps and detergents*: This has already been described in Section 4.3.

(vi) *Photographic plates and films*: Photographic plates or films are prepared by coating an emulsion of the light sensitive silver bromide in gelatin over glass plates or celluloid films.

(vii) *Rubber industry*: Latex is a colloidal solution of rubber particles which are negatively charged. Rubber is obtained by coagulation of latex.

(viii) *Industrial products*: Paints, inks, synthetic plastics, rubber, graphite lubricants, cement, etc., are all colloidal solutions.

Summary

Adsorption is the phenomenon of attracting and retaining the molecules of a substance on the surface of a solid resulting into a higher concentration on the surface than in the bulk. The substance adsorbed is known as **adsorbate** and the substance on which adsorption takes place is called **adsorbent.** In physisorption, adsorbate is held to the adsorbent by weak van der Waals forces, and in chemisorption, adsorbate is held to the adsorbent by strong chemical bond. Almost all solids adsorb gases. The extent of adsorption of a gas on a solid depends upon nature of gas, nature of solid, surface area of the solid, pressure of gas and temperature of gas. The relationship between the extent of adsorption (x/m) and pressure of the gas at constant temperature is known as **adsorption isotherm**.

A **catalyst** is a substance which enhances the rate of a chemical reaction without itself getting used up in the reaction. The phenomenon using catalyst is known as **catalysis**. In homogeneous catalysis, the catalyst is in the same phase as are the reactants, and in heterogeneous catalysis the catalyst is in a different phase from that of the reactants.

Colloidal solutions are intermediate between true solutions and suspensions. The size of the colloidal particles range from 1 to 1000 nm. A colloidal system consists of two phases - the dispersed phase and the dispersion medium. Colloidal systems are classified in three ways depending upon (i) physical states of the dispersed phase and dispersion medium (ii) nature of interaction between the dispersed phase and dispersion medium and (iii) nature of particles of dispersed phase. The colloidal systems show interesting optical, mechanical and electrical properties. The process of changing the colloidal particles in a sol into the insoluble precipitate by addition of some suitable electrolytes is known as **coagulation. Emulsions** are colloidal systems in which both dispersed phase and dispersion medium are liquids. These can be of: (i) **oil in water type** and (ii) **water in oil type.** The process of making emulsion is known as **emulsification.** To stabilise an emulsion, an emulsifying agent or emulsifier is added. Soaps and detergents are most frequently used as emulsifiers. Colloids find several applications in industry as well as in daily life.

IMPORTANT QUESTIONS

1. What is Catalysis? How is Catalysis classified? Give two examples for each type of Catalyses.
2. What are Emulsions? How are they classified? Describe the applications of Emulsions.
3. What is Adsorption? What are different types of adsorptions. Give any differences between Physical & Chemical Adsorption.
4. Define the terms
    a. Tyndall Effect
    b. Brownian Movement
    c. Emulsifying agent
    d. Promoters
    e. Poisons

CHAPTER 5

GENERAL PRINCIPLES OF METALLURGY

A few elements like carbon, sulphur, gold and noble gases, occur in free state while others in combined forms in the earth's crust. The extraction and isolation of an element from its combined form involves various principles of chemistry. A particular element may occur in a variety of compounds. The process of metallurgy and isolation should be such that it is chemically feasible and commercially viable. Still, some general principles are common to all the extraction processes of metals. For obtaining a particular metal, first we look for **minerals** which are naturally occurring chemical substances in the earth's crust obtainable by mining. Out of many minerals in which a metal may be found, only a few are viable to be used as sources of that metal. Such minerals are known as **ores**.

Rarely, an ore contains only a desired substance. It is usually contaminated with earthly or undesired materials known as **gangue**. The extraction and isolation of metals from ores involve the following major steps:

- Concentration of the ore,
- Isolation of the metal from its concentrated ore, and
- Purification of the metal.

The entire scientific and technological process used for isolation of the metal from its ores is known as **metallurgy**.

In the present Unit, first we shall describe various steps for effective concentration of ores. After that we shall discuss the principles of some of the common metallurgical processes. Those principles shall include the thermodynamic and electrochemical aspects involved in the effective reduction of the concentrated ore to the metal.

5.1 Occurrence of Metals

Elements vary in abundance. Among metals, aluminium is the most abundant. It is the third most abundant element in earth's crust (8.3% approx. by weight). It is a major component of many igneous minerals including mica and clays. Many gemstones are impure forms of Al_2O_3 and the impurities range from Cr (in 'ruby') to Co (in 'sapphire'). Iron is the second most abundant metal in the earth's crust. It forms a variety of compounds and their various uses make it a very important element. It is one of the essential elements in biological systems as well. The principal ores of aluminium, iron, copper and zinc have been given in Table 5.1.

Table 5.1: Principal Ores of Some Important Metals

Metal	Ores	Composition
Aluminium	Bauxite	$AlO_x(OH)_{3-2x}$ [where $0 < x < 1$]
	Kaolinite (a form of clay)	$[Al_2(OH)_4 Si_2O_5]$
Iron	Haematite	Fe_2O_3
	Magnetite	Fe_3O_4
	Siderite	$FeCO_3$
	Iron pyrites	FeS_2
Copper	Copper pyrites	$CuFeS_2$
	Malachite	$CuCO_3.Cu(OH)_2$
	Cuprite	Cu_2O
	Copper glance	Cu_2S
Zinc	Zinc blende or Sphalerite	ZnS
	Calamine	$ZnCO_3$
	Zincite	ZnO

For the purpose of extraction, bauxite is chosen for aluminium. For iron, usually the oxide ores which are abundant and do not produce polluting gases (like SO_2 that is produced in case iron pyrites) are taken. For copper and zinc, any of the listed ores (Table 5.1) may be used depending upon availability and other relevant factors. Before proceeding for concentration, ores are graded and crushed to reasonable size.

5.2 Concentration of Ores

Removal of the unwanted materials (e.g., sand, clays, etc.) from the ore is known as *concentration, dressing* or *benefaction.* It involves several steps and selection of these steps depends upon the differences in physical properties of the compound of the metal present and that of the *gangue.* The type of the metal, the available facilities and the environmental factors are also taken into consideration. Some of the important procedures are described below.

Hydraulic Washing or Legivation

This is based on the differences in gravities of the ore and the *gangue* particles. It is therefore a type of *gravity separation*. In one such process, an upward stream of running water is used to wash the powdered ore. The lighter gangue particles are washed away and the heavier ores are left behind.

Magnetic Separation

This is based on differences in magnetic properties of the ore components. If either the ore or the gangue (one of these two) is capable of being attracted by a magnetic field, then such separations are carried out (e.g., in case of iron ores). The ground ore is carried on a conveyer belt which passes over a

magnetic roller Fig.5.1.

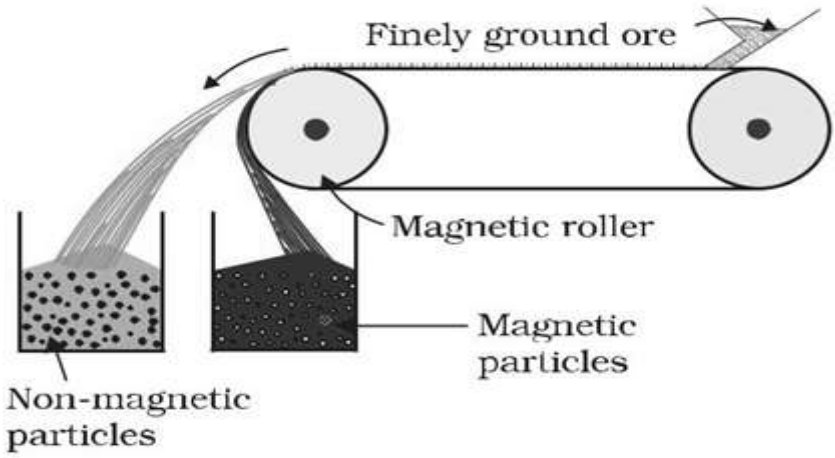


Fig 5.1 Magnetic Separation (Schematic)

Froth Floatation Method

This method has been in use for removing gangue from sulphide ores. In this process, a suspension of the powdered ore is made with water. To it, *collectors* and *froth stabilisers* are added. Collectors (e. g., pine oils, fatty acids, xanthates, etc.) enhance non-wettability of the mineral particles and froth stabilizers (e. g., cresols, aniline) stabilise the froth.

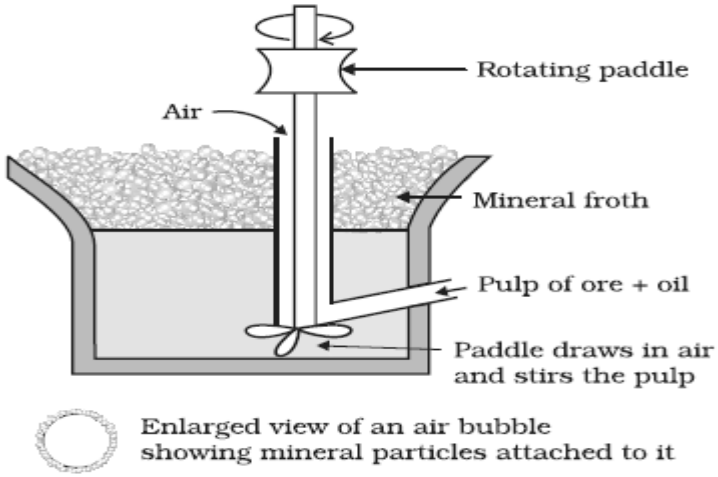


Fig 5.2 Froth flotation process (schematic)

The mineral particles become wet by oils while the gangue particles by water. A rotating paddle agitates the mixture and draws air in it. As a result, froth is formed which carries the mineral particles. The froth is light and is skimmed off. It is then dried for recovery of the ore particles.

Sometimes, it is possible to separate two sulphide ores by adjusting proportion of oil to water or by using '*depressants*'. For example, in case of an ore containing ZnS and PbS, the depressant used is NaCN. It selectively prevents ZnS from coming to the froth but allows PbS to come with the froth.

Leaching

Leaching is often used if the ore is soluble in some suitable solvent.

(a) Leaching of alumina from bauxite

The principal ore of aluminium, bauxite, usually contains SiO_2, iron oxides and titanium oxide (TiO_2) as impurities. Concentration is carried out by digesting the powdered ore with a concentrated solution of NaOH at 473 – 523 K and 35 – 36 bar pressure. This way, Al_2O_3 is leached out as sodium aluminate (and SiO_2 too as sodium silicate) leaving the impurities behind:

$$Al_2O_3(s) + 2NaOH(aq) + 3H_2O(l) \rightarrow 2Na[Al(OH)_4](aq) \quad (5.1)$$

The aluminate in solution is neutralised by passing CO_2 gas and hydrated Al_2O_3 is precipitated. At this stage, the solution is seeded with freshly prepared samples of hydrated Al_2O_3 which induces the precipitation:

$$2Na[Al(OH)_4](aq) + CO_2(g) \rightarrow Al_2O_3.xH_2O(s) + 2NaHCO_3(aq) \quad (5.2)$$

The sodium silicate remains in the solution and hydrated alumina is filtered, dried and heated to give back pure Al_2O_3:

$$Al_2O_3.xH_2O(s) \xrightarrow{1470\ K} Al_2O_3(s) + xH_2O(g) \quad (5.3)$$

(b) Other examples

In the metallurgy of silver and that of gold, the respective metal is leached with a dilute solution of NaCN or KCN in the presence of air (for O_2) from which the metal is obtained later by replacement:

$$4M(s) + 8CN^-(aq) + 2H_2O(aq) + O_2(g) \rightarrow 4[M(CN)_2]^-(aq) + 4OH^-(aq) \quad (M = Ag \text{ or } Au) \quad (6.4)$$

$$2[M(CN)_2]^-(aq) + Zn(s) \rightarrow [Zn(CN)_4]^{2-}(aq) + 2M(s) \quad (6.5)$$

5.3 Extraction of Crude Metal from Concentrated Ore

The concentrated ore must be converted into a form which is suitable for reduction. Usually the sulphide ore is converted to oxide before reduction. Oxides are easier to reduce (for the reason see box). Thus isolation of metals from concentrated ore involves two major steps *viz*., **i) conversion to oxide, and ii) reduction of the oxide to metal.**

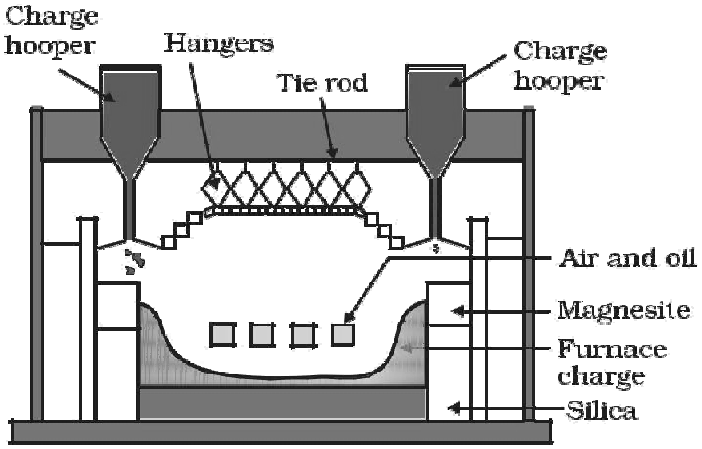


Fig. 5.3 A section of a modern reverberatory furnace

(a) Conversion to oxide

(i) Calcination

Calcinaton involves heating when the volatile matter escapes leaving behind the metal oxide:

$$Fe_2O_3.xH_2O(s) \xrightarrow{\Delta} Fe_2O_3\ (s) + xH_2O(g)$$
$$ZnCO_3\ (s) \xrightarrow{\Delta} ZnO(s) + CO_2(g)$$
$$CaCO_3.MgCO_3(s) \xrightarrow{\Delta} CaO(s) + MgO(s\ ) + 2CO_2(g)$$

(ii) Roasting

In roasting, the ore is heated in a regular supply of air in a furnace at a temperature below the melting point of the metal. Some of the reactions involving sulphide ores are:

$$2ZnS + 3O_2 \rightarrow 2ZnO + 2SO_2$$
$$2PbS + 3O_2 \rightarrow 2PbO + 2SO_2$$
$$2Cu_2S + 3O_2 \rightarrow 2Cu_2O + 2SO_2$$

The sulphide ores of copper are heated in reverberatory furnace. If the ore contains iron, it is mixed with silica before heating. Iron oxide 'slags of '***** as iron silicate and copper is produced in the form of *copper matte* which contains Cu_2S and FeS.

$$FeO + SiO_2 \rightarrow \underset{\text{(slag)}}{FeSiO_3}$$

The SO_2 produced is utilised for manufacturing H_2SO_4 .

(b) Reduction of oxide to the metal

Reduction of the metal oxide usually involves heating it with some other substance acting as a reducing agent (C or CO or even another metal). The reducing agent (e.g., carbon) combines with the oxygen of the metal oxide.

$$M_xO_y + yC \rightarrow xM + y\ CO$$

Some metal oxides get reduced easily while others are very difficult to be reduced (reduction means electron gain or electronation). In any case, heating is required. To understand the variation in the temperature requirement for thermal reductions *(pyrometallurgy)* and to predict which element will suit as the reducing agent for a given metal oxide (M_xO_y), Gibbs energy interpretations are made.

5.4 Extraction of iron from its oxides

Oxide ores of iron, after concentration through calcination/roasting (to remove water, to decompose carbonates and to oxidise sulphides) are mixed with limestone and coke and fed into a *Blast furnace* from its top. Here, the oxide is reduced to the metal. Thermodynamics helps us to understand how coke reduces the oxide and why this furnace is chosen. One of the main reduction steps in this process is:

$$FeO(s) + C(s) \rightarrow Fe(s/l) + CO\ (g)$$

It can be seen as a couple of two simpler reactions. In one, the reduction of FeO is taking place and in the other, C is being oxidised to CO:

$$FeO(s) \rightarrow Fe(s) + \frac{1}{2}O_2(g) \qquad [\Delta G_{(FeO,\ Fe)}]$$

$$C(s) + \frac{1}{2}O_2\ (g) \rightarrow CO\ (g) \qquad [\Delta G_{(C,\ CO)}]$$

When both the reactions take place to yield the equation (6.23), the net Gibbs energy change becomes:

$$\Delta G_{(C,\ CO)} + \Delta G_{(FeO,\ Fe)} = \Delta_r G$$

Naturally, the resultant reaction will take place when the right hand side in equation 6.27 is negative. In G^0 vs T plot representing reaction 6.25, the plot goes upward and that representing the change C→CO

(C,CO) goes downward. At temperatures above 1073K (approx.), the

C,CO line comes below the Fe,FeO line [$G_{(C,\ CO)} < G_{(Fe,\ FeO)}$]. So in this range, coke will be reducing the FeO and will itself be oxidised to CO.

In a similar way the reduction of Fe_3O_4 and Fe_2O_3 at relatively lower temperatures by CO can be explained on the basis of lower lying points of intersection of their curves with the CO, CO_2 curve in Fig. 5.4.

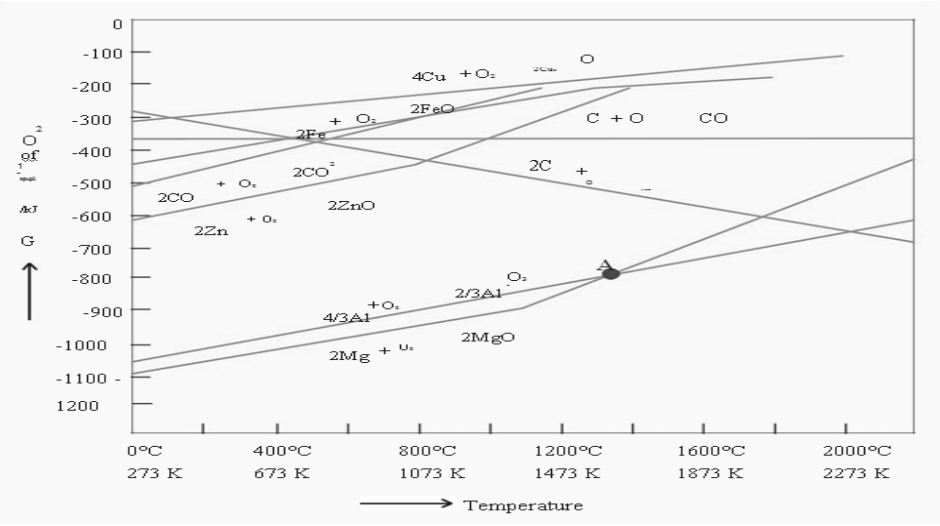


Fig. 5.4 Gibbs energy (G^V) vs T plots (schematic) for formation of some oxides (Ellingham diagram)

In the Blast furnace, reduction of iron oxides takes place in different temperature ranges. Hot air is blown from the bottom of the furnace and coke is burnt to give temperature upto about 2200K in the lower portion itself. The burning of coke therefore supplies most of the heat required in the process. The CO and heat moves to upper part of the furnace. In upper part, the temperature is lower and the iron oxides (Fe_2O_3 and Fe_3O_4) coming from the top are reduced in steps to FeO. Thus, the reduction reactions taking place in the lower temperature range and in the higher temperature range, depend on the points of corresponding intersections in the $_rG^0$ vs T plots. These reactions can be summarised as follows:

At 500 – 800 K (lower temperature range in the blast furnace) –

$$3\,Fe_2O_3 + CO \rightarrow 2\,Fe_3O_4 + CO_2$$

$$Fe_3O_4 + 4\,CO \rightarrow 3Fe + 4\,CO_2$$

$$Fe_2O_3 + CO \rightarrow 2FeO + CO_2$$

At 900 – 1500 K (higher temperature range in the blast furnace:

$$C + CO_2 \rightarrow 2\,CO$$

$$FeO + CO \rightarrow Fe + CO_2$$

Limestone is also decomposed to CaO which removes silicate impurity of the ore as slag. The slag is in molten state and separates out from iron.

The iron obtained from Blast furnace contains about 4% carbon and many impurities in smaller amount (e.g., S, P, Si, Mn). This is known as *pig iron* and cast into variety of shapes. *Cast iron* is different from *pig iron* and is made by melting pig iron with scrap iron and coke using hot air blast. It has slightly lower carbon content (about 3%) and is extremely hard and brittle.

Further Reductions

Wrought iron or malleable iron is the purest form of commercial iron and is prepared from cast iron by oxidising impurities in a reverberatory furnace lined with haematite. This haematite oxidises carbon to carbon monoxide:

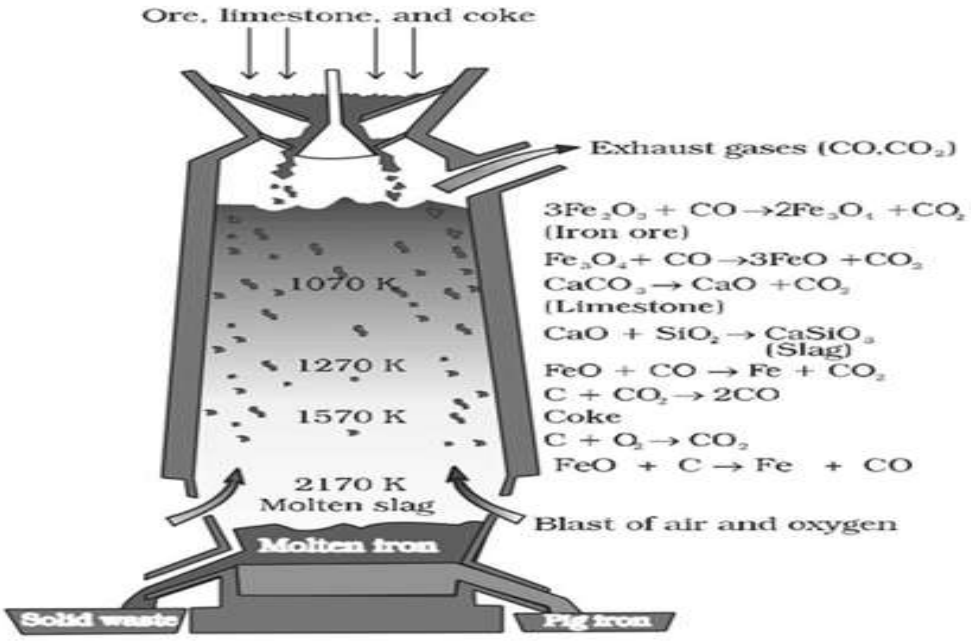


Fig 5.5 Blast Furnace

$$Fe_2O_3 + 3\,C \rightarrow 2\,Fe + 3\,CO$$

Limestone is added as a flux and sulphur, silicon and phosphorus are oxidised and passed into the slag. The metal is removed and freed from the slag by passing through rollers.

(a) Extraction of copper from cuprous oxide [copper(I) oxide]

In the graph of ΔrG0 vs T for formation of oxides (Fig. 6.4), the Cu2O line is almost at the top. So it is quite easy to reduce oxide ores of copper directly to the metal by heating with coke (both the lines of C, CO and C, CO2 are at much lower positions in the graph particularly after 500 – 600K). However most of the ores are sulphide and some may also contain iron. The sulphide ores are roasted/smelted to give oxides:

$$2Cu_2S + 3O_2 \rightarrow 2Cu_2O + 2SO_2$$

The oxide can then be easily reduced to metallic copper using coke:

$$Cu_2O + C \rightarrow 2\,Cu + CO$$

Limestone is added as a flux and sulphur, silicon and phosphorus are oxidised and passed into the slag. The metal is removed and freed from the slag by passing through rollers.

(b) Extraction of copper from cuprous oxide [copper(I) oxide]

In the graph of ${}_rG^0$ vs T for formation of oxides (Fig. 6.4), the Cu_2O line is almost at the top. So it is quite easy to reduce oxide ores of copper directly to the metal by heating with coke (both the lines of C, CO and C, CO_2 are at much lower positions in the graph particularly after 500 – 600K). However most of the ores are sulphide and some may also contain iron. The sulphide ores are roasted/smelted to give oxides:

$$2Cu_2S + 3O_2 \rightarrow 2Cu_2O + 2SO_2$$

The oxide can then be easily reduced to metallic copper using coke:

$$Cu_2O + C \rightarrow 2\,Cu + CO$$

In actual process, the ore is heated in a reverberatory furnace after mixing with silica. In the furnace, iron oxide 'slags of' as iron silicate and copper is produced in the form of *copper matte*. This contains Cu_2S and FeS.

$$FeO + SiO_2 \rightarrow \underset{\text{(Slag)}}{FeSiO_3}$$

Copper matte is then charged into silica lined convertor. Some silica is also added and hot air blast is blown to convert the remaining FeS_2, FeO and Cu_2S/Cu_2O to the metallic copper. Following reactions take place:

$$2FeS + 3O_2 \rightarrow 2FeO + 2SO_2$$
$$FeO + SiO_2 \rightarrow FeSiO_3$$
$$2Cu_2S + 3O_2 \rightarrow 2Cu_2O + 2SO_2$$
$$2Cu_2O + Cu_2S \rightarrow 6Cu + SO_2$$

The solidified copper obtained has blistered appearance due to the evolution of SO_2 and so it is called *blister copper.*

(c) Extraction of zinc from zinc oxide

The reduction of zinc oxide is done using coke. The temperature in this case is higher than that in case of copper. For the purpose of heating, the oxide is made into brickettes with coke and clay.

$$ZnO + C \xrightarrow{\text{coke, 673 K}} Zn + CO$$

The metal is distilled off and collected by rapid chilling.

Extraction of Iron, Copper and Zinc from their Ores:

EXTRATION OF IRON

Haematite Fe_2O_3

Concentrated ore ↓ **Crushed Ore is washed with stream of water**

Calcine d ore ↓ **Roasting in air/ Calcination Moisture and volatile impurities are eliminated FeO changes to Fe_2O_3**

Pig Iron ↓ **Calcined ore, coke and limestone in 8:4:1, reduced with CO in blast furnace. Sio_2 impurity eliminated as slag, $CaSiO_3$**

EXTRATION OF COPPER

Copper pyrites $CuFeS_2$

Concentrated Ore ↓ **Froth floatation**

Roasted Ore ↓ **Roasting (moisture and volatile impurities eliminated) Formation of Cu_2S and FeS**

Matte Cu_2S and FeS ↓ **Mixed with coke and sand (flux), subjected to smelting in blast furnace FeS is partly eliminated as $FeSiO_3$ (Slag)**

Blister Copper

Pure Copper Metal ↓ **Hot air and sand (SiO_2)**
FeS is removed as $FeSiO_3$ (Slag)
Cu_2O and Cu_2S react to give copper (self reduction)

↓ **Refining poling and eletrolysis**

EXTRATION OF ZINC

Zinc Blende ZnS

Concentrated ore ↓ **Concentration by forth floatation**

ZnO ↓ **Roasting in air**

Zinc ↓ **Reduction with coke at 1673K**

Pure Zinc Metal ↓ **Purification by fractional distillation / electrolysis**

5.5 Refining of Crude Metal

A metal extracted by any method is usually contaminated with some impurity. For obtaining metals of high purity, several techniques are used depending upon the differences in properties of the metal and the impurity. Some of them are listed below.

(a) Distillation **(b)** Liquation

(c) Electrolysis **(d)** Zone refining

These are described in detail here.

Distillation

This is very useful for low boiling metals like zinc and mercury. The impure metal is evaporated to obtain the pure metal as distillate.

Liquation

In this method a low melting metal like tin can be made to flow on a sloping surface. In this way it is separated from higher melting impurities.

Electrolytic refining

In this method, the impure metal is made to act as anode. A strip of the same metal in pure form is used as cathode. They are put in a suitable electrolytic bath containing soluble salt of the same metal. The more basic metal remains in the solution and the less basic ones go to the anode mud. This process is also explained using the concept of electrode potential, over potential, and Gibbs energy which you have seen in previous sections. The reactions are:

$$\text{Anode: } M \rightarrow M^{n+} + ne^{-}$$

$$\text{Cathode: } M^{n+} + ne^{-} \rightarrow M$$

Copper is refined using an electrolytic method. Anodes are of impure copper and pure copper strips are taken as cathode. The electrolyte is acidified solution of copper sulphate and the net result of electrolysis is the transfer of copper in pure form from the anode to the cathode:

$$\text{Anode: } Cu \rightarrow Cu^{2+} + 2e^{-}$$

$$\text{Cathode: } Cu^{2+} + 2e^{-} \rightarrow Cu$$

Impurities from the blister copper deposit as anode mud which contains antimony, selenium, tellurium, silver, gold and platinum; recovery of these elements may meet the cost of refining.

Zinc may also be refined this way.

(d) Zone refining

This method is based on the principle that the impurities are more soluble in the melt than in the solid state of the metal. A circular mobile heater is fixed at one end of a rod of the impure metal (Fig. 5.7). The molten zone moves along with the heater which is moved forward. As the heater moves forward, the pure metal crystallises out of the melt and the impurities pass on into the adjacent molten zone. The process is repeated several times and the heater is moved in the same direction. At one end, impurities get concentrated. This end is cut off. This method is very useful for producing semiconductor and other metals of very high purity, e.g., germanium, silicon, boron, gallium and indium.

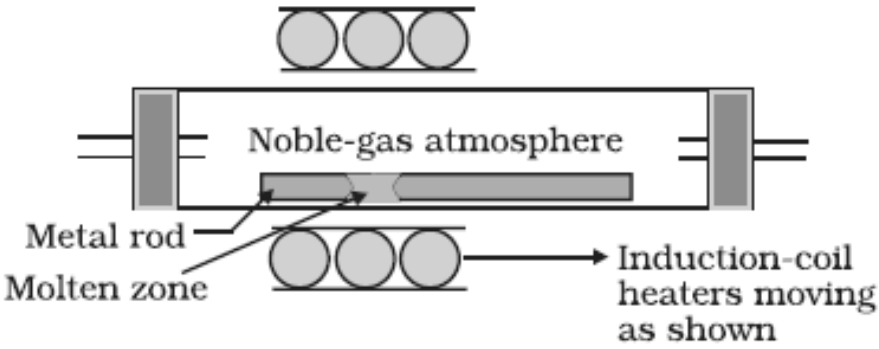


Fig 5.7 Zone refining process

Uses of Aluminium, Copper, Zinc and Iron

Aluminium foils are used as wrappers for chocolates. The fine dust of the metal is used in paints and lacquers. Aluminium, being highly reactive, is also used in the extraction of chromium and manganese from their oxides. Wires of aluminium are used as electricity conductors. Alloys containing aluminium, being light, are very useful.

Copper is used for making wires used in electrical industry and for water and steam pipes. It is also used in several alloys that are rather tougher than the metal itself, e.g., brass (with zinc), bronze (with tin) and coinage alloy (with nickel).

Zinc is used for galvanising iron. It is also used in large quantities in batteries, as a constituent of many alloys, e.g., brass, (Cu 60%, Zn 40%) and german silver (Cu 25-30%, Zn 25-30%, Ni 40–50%). Zinc dust is used as a reducing agent in the manufacture of dye-stuffs, paints, etc.

Cast iron, which is the most important form of iron, is used for casting stoves, railway sleepers, gutter pipes , toys, etc. It is used in the manufacture of wrought iron and steel. Wrought iron is used in making anchors, wires, bolts, chains and agricultural implements. Steel finds a number of uses. Alloy steel is obtained when other metals are added to it. Nickel steel is used for making cables, automobiles and aeroplane parts, pendulum, measuring tapes, chrome steel for cutting tools and crushing machines, and stainless steel for cycles, automobiles, utensils, pens, etc.

Summary

Metals are required for a variety of purposes. For this, we need their extraction from the minerals in which they are present and from which their extraction is commercially feasible.These minerals are known as **ores**. Ores of the metal are associated with many impurities. Removal of these impurities to certain extent is achieved in **concentration** steps. The concentrated ore is then treated chemically for obtaining the metal. Usually the metal compounds (e.g., oxides, sulphides) are reduced to the metal. The reducing agents used are carbon, CO or even some metals.

In these reduction processes, the **thermodynamic** and **electrochemical** concepts are given due consideration. The metal oxide reacts with a reducing agent; the oxide is reduced to the metal and the reducing agent is oxidised. In the two reactions, the net Gibbs energy change is negative, which becomes more negative on raising the temperature. Conversion of the physical states from solid to liquid or to gas, and formation of gaseous states favours decrease in the Gibbs energy for the entire system. This concept is graphically displayed in plots of G^0 vs T (Ellingham diagram) for such oxidation/reduction reactions at different temperatures. The concept of electrode potential is useful in the isolation of metals (e.g., Al, Ag, Au) where the sum of the two redox couples is +ve so that the Gibbs energy change is negative. The metals obtained by usual methods still contain minor impurities. Getting pure metals require **refining**.

Refining process depends upon the differences in properties of the metal and the impurities. Extraction of aluminium is usually carried out from its bauxite ore by leaching it with NaOH. Sodium aluminate, thus formed, is separated and then neutralised to give back the hydrated oxide, which is then electrolysed using cryolite as a flux. Extraction of iron is done by reduction of its oxide ore in blast furnace. Copper is extracted by smelting and heating in a reverberatory furnace. Extraction of zinc from zinc oxides is done using coke. Several methods are employed in refining the metal. Metals, in general, are very widely used and have contributed significantly in the development of a variety of industries.

A Summary of the Occurrence and Extraction of some Metals is Presented in the following Table

Metal	Occurrence	Common method of extraction	Remarks
Aluminium	Bauxite, Al_2O_3. *x* H_2O Cryolite, Na_3AlF_6	Electrolysis of Al_2O_3 dissolved in molten Na_3AlF_6	For the extraction, a good source of electricity is required.
Iron	Haematite, Fe_2O_3 Magnetite, Fe_3O_4	Reduction of the oxide with CO and coke in Blast furnace	Temperature approaching 2170 K is required.
Copper	Copper pyrites, $CuFeS_2$ Copper glance, Cu_2S Malachite, $CuCO_3.Cu(OH)_2$ Cuprite, Cu_2O	Roasting of sulphide partially and reduction	It is self reduction in a specially designed converter. The reduction takes place easily. Sulphuric acid leaching is also used in hydrometallurgy from low grade ores
Zinc	Zinc blende or Sphalerite, ZnS Calamine, $ZnCO_3$ Zincite, ZnO	Roasting followed by reduction with coke	The metal may be purified by fractional distillation.

IMPORTANT QUESTIONS

1. Differentiate Roasting and calcinations with suitable examples
2. Explain Zone Refining and Poling
3. Give the composition of the following alloys
    a. Brass
    b. Bronze
    c. German Sliver
4. Explain the terms Gangue, Slag, Matte, Blister Copper and Flux.
5. Explain the purification of Sulphide Ore by Froth Floatation method

CHAPTER 6

p-BLOCK ELEMENTS (Group – 15 elements)

Group 15 includes nitrogen, phosphorus, arsenic, antimony and bismuth. As we go down the group, there is a shift from non-metallic to metallic through metalloidic character. Nitrogen and phosphorusare non-metals, arsenic and antimony metalloids and bismuth is a typical metal.

6.1 Occurrence

Molecular nitrogen comprises 78% by volume of the atmosphere. In the earth's crust, it occurs as sodium nitrate, $NaNO_3$ (called Chile saltpetre) and potassium nitrate (Indian salt petre). It is found in the form of proteins in plants and animals. Phosphorus occurs in minerals of the apatite family, $Ca_9(PO_4)_6$. CaX_2 (X = F, Cl or OH) (e.g., fluorapatite $Ca_9(PO_4)_6$. CaF_2 which are the main components of phosphate rocks.

Phosphorus is an essential constituent of animal and plant matter. It is present in bones as well as in living cells. Phosphoproteins are present in milk and eggs. Arsenic, antimony and bismuth are found mainly as sulphide minerals.

Electronic Configuration

The valence shell electronic configuration of these elements is ns^2np^3. The *s* orbital in these elements is completely filled and *p* orbitals are half-filled, making their electronic configuration extra stable.

Atomic and Ionic Radii

Covalent and ionic (in a particular state) radii increase in size down the group. There is a considerable increase in covalent radius from N to P. However, from As to Bi only a small increase in covalent radius is observed. This is due to the presence of completely filled *d* and/or *f* orbitals in heavier members.

Ionisation Enthalpy

Ionisation enthalpy decreases down the group due to gradual increase in atomic size. Because of the extra stable half-filled *p* orbitals electronic configuration and smaller size, the ionisation enthalpy of the group 15 elements is much greater than that of group 14 elements in the corresponding periods.

Electro negativity

The electro negativity value, in general, decreases down the group with increasing atomic size. However, amongst the heavier elements, the difference is not that much pronounced.

Physical Properties

All the elements of this group are polyatomic. Dinitrogen is a diatomic gas while all others are solids. Metallic character increases down the group. Nitrogen and phosphorus are non-metals, arsenic and antimony metalloids and bismuth is a metal. This is due to decrease in ionisation enthalpy and increase in atomic size. The boiling points, in general, increase from top to bottom in the group but the melting point increases upto arsenic and then decreases upto bismuth. Except nitrogen, all the elements show allotropy.

Chemical Properties

Oxidation states and trends in chemical reactivity

The common oxidation states of these elements are –3, +3 and +5 The tendency to exhibit –3 oxidation state decreases down the group duento increase in size and metallic

character. In fact last member of the group, bismuth hardly forms any compound in –3 oxidation state. The stability of +5 oxidation state decreases down the group. The only well characterized Bi (V) compound is BiF_5. The stability of +5 oxidation state decreases and that of +3 state increases (due to inert pair effect) down the group. Nitrogen exhibits + 1, + 2, + 4 oxidation states also when it reacts with oxygen. Phosphorus also shows +1 and +4 oxidation states in some oxoacids. In the case of nitrogen, all oxidation states from +1 to +4 tend to disproportionate in acid solution. For example,

$$3HNO_2 \rightarrow HNO_3 + H_2O + 2NO$$

Similarly, in case of phosphorus nearly all intermediate oxidation states disproportionate into +5 and –3 both in alkali and acid. However +3 oxidation state in case of arsenic, antimony and bismuth become increasingly stable with respect to disproportionation. Nitrogen is restricted to a maximum covalency of 4 since only four (one *s* and three *p*) orbitals are available for bonding. The heavier elements have vacant *d* orbitals in the outermost shell which can be used for bonding (covalency) and hence, expand their covalence as in PF_6

Anomalous properties of nitrogen

Nitrogen differs from the rest of the members of this group due to its smaller size, high electronegativity, high ionisation enthalpy and non-availability of *d* orbitals. Nitrogen has unique ability to form ***pπ -pπ* multiple** bonds with itself and with other elements having small size and high electronegativity (e.g., C, O). Heavier elements of this group do not form *pπ -pπ* bonds as their atomic orbitals are so large and diffuse that they cannot have effective overlapping. Thus, nitrogen exists as a diatomic molecule with a triple bond (one *s* and two *p*) between the two atoms. Consequently, its bond enthalpy (941.4 kJ mol–1) is very high. On the contrary, phosphorus, arsenic and antimony form single bonds as P–P, As–As and Sb–Sb while bismuth forms metallic bonds in elemental state. However, the single N–N bond is weaker than the single P–P bond because of high interelectronic repulsion of the non-bonding electrons, owing to the small bond length. As a result the catenation tendency is weaker in nitrogen. Another factor which affects the chemistry of nitrogen is the absence of *d* orbitals in its valence shell. Besides restricting its covalency to four, nitrogen cannot form ***dπ –pπ* bond** as the heavier elements can e.g., $R_3P = O$ or $R_3P = CH_2$ (R = alkyl group). Phosphorus and arsenic can form ***dπ –dπ* bond** also with transition metals when their compounds like $P(C_2H_5)_3$ and $As(C_6H_5)_3$ act as ligands.

(i) Reactivity towards hydrogen

All the elements of Group 15 form hydrides of the type EH_3 where E = N, P, As, Sb or Bi. The hydrides show regular gradation in their properties. The stability of hydrides decreases from NH_3 to BiH_3 which can be observed from their bond dissociation enthalpy. Consequently, the reducing character of the hydrides increases. Ammonia is only a mild reducing agent while BiH_3 is the strongest reducing agent amongst all the hydrides. Basicity also decreases in the order $NH_3 > PH_3 > AsH_3 > SbH_3 > BiH_3$.

(ii) Reactivity towards oxygen

All these elements form two types of oxides: $E2O_3$ and E_2O_5. The oxide in the higher oxidation state of the element is more acidic than that of lower oxidation state.Their acidic character decreases down the group. The oxides of the type E_2O_3 of nitrogen and phosphorus are purely acidic, that of arsenic and antimony amphoteric and those of bismuth is predominantly basic.

(iii) Reactivity towards halogens

These elements react to form two series of halides: EX_3 and EX_5. Nitrogen does not form pentahalide due to non-availability of the *d* orbitals in its valence shell. Pentahalides are more covalent than trihalides. All the trihalides of these elements except those of nitrogen are stable. In case of nitrogen, only NF_3 is known to be stable. Trihalides except BiF3 are predominantly covalent in nature.

(iv) Reactivity towards metals

All these elements react with metals to form their binary compounds exhibiting –3 oxidation state, such as, Ca_3N_2 (calcium nitride) Ca_3P_2 (calcium phosphide), Na_3As_2 (sodium arsenide), Zn_3Sb_2 (zinc antimonide) and Mg_3Bi_2 (magnesium bismuthide).

6.2 Dinitrogen

Preparation

Dinitrogen is produced commercially by the liquefaction and fractional distillation of air. Liquid dinitrogen (b.p. 77.2 K) distils out first leaving behind liquid oxygen (b.p. 90 K). In the laboratory, dinitrogen is prepared by treating an aqueous solution of ammonium chloride with sodium nitrite.

$$NH_4Cl(aq) + NaNO_2(aq) \rightarrow N_2(g) + 2H_2O(l) + NaCl\ (aq)$$

Small amounts of NO and HNO3 are also formed in this reaction,these impurities can be removed by passing the gas through aqueous sulphuric acid containing potassium dichromate. It can also be obtained by the thermal decomposition of ammonium dichromate.

$$(NH_4)_2Cr_2O_7 \xrightarrow{Heat} N_2 + 4H_2O + Cr_2O_3$$

Very pure nitrogen can be obtained by the thermal decomposition of sodium or barium azide.

$$Ba(N_3)_2 \rightarrow Ba + 3N_2$$

Properties

Dinitrogen is a colourless, odourless, tasteless and non-toxic gas. It has two stable isotopes: $_{14}N$ and $_{15}N$. It has a very low solubility in water (23.2 cm^3 per litre of water at 273 K and 1 bar pressure) and low freezing and boiling points. Dinitrogen is rather inert at room temperature because of the high bond enthalpy of N≡N bond. Reactivity, however, increases rapidly with rise in temperature. At higher temperatures, it directly combines with some metals to form predominantly ionic nitrides and with non-metals, covalent nitrides. A few typical reactions are:

$$6Li + N_2 \xrightarrow{Heat} 2Li_3N$$
$$3Mg + N_2 \xrightarrow{Heat} Mg_3N_2$$

It combines with hydrogen at about 773 K in the presence of a catalyst (Haber's Process) to form ammonia:

$$N_2(g) + 3H_2(g) \underset{}{\overset{773\ k}{\rightleftharpoons}} 2NH_3(g); \qquad \Delta_f H^{\ominus} = -46.1\ kJmol^{-1}$$

Dinitrogen combines with dioxygen only at very high temperature (at about 2000 K) to form nitric oxide, NO.

$$N_2 + O_2(g) \underset{}{\overset{Heat}{\rightleftharpoons}} 2NO(g)$$

Uses

The main use of dinitrogen is in the manufacture of ammonia and other industrial chemicals containing nitrogen, (e.g., calcium cyanamide). It also finds use where an inert atmosphere is required (e.g., in iron and steel industry, inert diluent for reactive chemicals). Liquid dinitrogen is used as a refrigerant to preserve biological materials, food items and in cryosurgery.

6.3 Compounds of Nitrogen

Preparation of Ammonia:

Ammonia is present in small quantities in air and soil where it is formed by the decay of nitrogenous organic matter e.g., urea.

$$NH_2CONH_2 + 2H_2O \rightarrow (NH_4)_2CO_3 \rightleftharpoons 2NH_3 + H_2O + CO_2$$

On a small scale ammonia is obtained from ammonium salts which decompose when treated with caustic soda or lime.

$$2NH_4Cl + Ca(OH)_2 \rightarrow 2NH_3 + 2H_2O + CaCl_2$$

$$(NH_4)_2 SO_4 + 2NaOH \rightarrow 2NH_3 + 2H_2O + Na_2SO_4$$

On a large scale, ammonia is manufactured by Haber's process.

$$N_2(g) + 3H_2(g) \rightleftharpoons 2NH_3(g); \qquad \Delta_f H^{\ominus} = -46.1 \text{ kJ mol}^{-1}$$

In accordance with Le Chatelier's principle, high pressure would favour the formation of ammonia. The optimum conditions for the production of ammonia are a pressure of 200 × 10 5Pa (about 200 atm), a temperature of ~ 700 K and the use of a catalyst such as iron oxide with small amounts of K_2O and Al_2O_3 to increase the rate of attainment of equilibrium

Properties

Ammonia is a colourless gas with a pungent odour. Its freezing and boiling points are 198.4 and 239.7 K respectively. In the solid and liquid states, it is associated through hydrogen bonds as in the case of water and that accounts for its higher melting and boiling points than expected on the basis of its molecular mass. The ammonia molecule is trigonal pyramidal with the nitrogen atom at the apex. It has three bond pairs and one lone pair of electrons as shown in the structure. Ammonia gas is highly soluble in water. Its aqueous solution is weakly basic due to the formation of OH– ions.

$$NH_3(g) + H_2O(l) \rightleftharpoons NH_4^+(aq) + OH^-(aq)$$

It forms ammonium salts with acids, e.g., NH_4Cl, $(NH_4)_2 SO_4$, etc. As a weak base, it precipitates the hydroxides of many metals from their salt solutions. For example,

$$2FeCl_3(aq) + 3NH_4OH(aq) \rightarrow Fe_2O_3.xH_2O(s) + 3NH_4Cl(aq)$$

(brown ppt)

$$ZnSO_4(aq) + 2NH_4OH(aq) \rightarrow Zn(OH)_2(s) + (NH_4)_2 SO_4(aq)$$

(white ppt)

The presence of a lone pair of electrons on the nitrogen atom of the ammonia molecule makes it a Lewis base. It donates the electron pair and forms linkage with metal ions and the formation of such complex compounds finds applications in detection of metal ions such as Cu^{2+}, Ag+:

$$Cu^{2+}(aq) + 4\,NH_3(aq) \rightleftharpoons [Cu(NH_3)_4]^{2+}(aq)$$
(blue) (deep blue)

$$Ag^{+}(aq) + Cl^{-}(aq) \rightarrow AgCl(s)$$
(colourless) (white ppt)

$$AgCl(s) + 2NH_3(aq) \rightarrow [Ag(NH_3)_2]Cl(aq)$$
(white ppt) (colourless)

Uses

Ammonia is used to produce various nitrogenous fertilizers (ammonium nitrate, urea, ammonium phosphate and ammonium sulphate) and in the manufacture of some inorganic nitrogen compounds, the most important one being nitric acid. Liquid ammonia is also used as a refrigerant.

6.4 Oxides of Nitrogen

Nitrogen forms a number of oxides in different oxidation states. The names, formulas, preparation and physical appearance of these oxides are given in the following table.

Oxides of Nitrogen

Name	Formula	Oxidation state of nitrogen	Common methods of preparation	Physical appearance and chemical nature
Dinitrogen oxide [Nitrogen (I) oxide]	N_2O	+1	$NH_4NO_3 \xrightarrow{\text{Heat}} N_2O + 2H_2O$	Colourless gas neutral
Nitrogen monoxide [Nitrogen(II) oxide]	NO	+2	$2NaNO_2 + 2FeSO_4 + 3H_2SO_4 \rightarrow Fe_2(SO_4)_3 + 2NaHSO_4 + 2H_2O + 2NO$	Colourless gas neutral
Dinitrogen trioxide [Nitrogen(III) oxide]	N_2O_3	+3	$2NO + N_2O_4 \xrightarrow{250K} 2N_2O_3$	Blue solid acidic
Nitrogen dioxide	NO_2	+4	$2Pb(NO_3)_2 \xrightarrow{673K} 4NO_2 + 2PbO$	Brown gas acidic
Dinitrogen tetraoxide [Nitrogen(IV) oxide]	N_2O_4	+4	$2NO_2 \underset{\text{Heat}}{\overset{\text{Cool}}{\rightleftharpoons}} N_2O_4$	Colour solid/liquid.gas
Dinitrogen pentaoxide [Nitrogen(V) oxide]	N_2O_5	+5	$4HNO_3 + P_4O_{10} \rightarrow 4HPO_3 + 2N_2O_5$	Colour solid. acidic

Structures of Oxides of Nitrogen

6.5 Nitric Acid

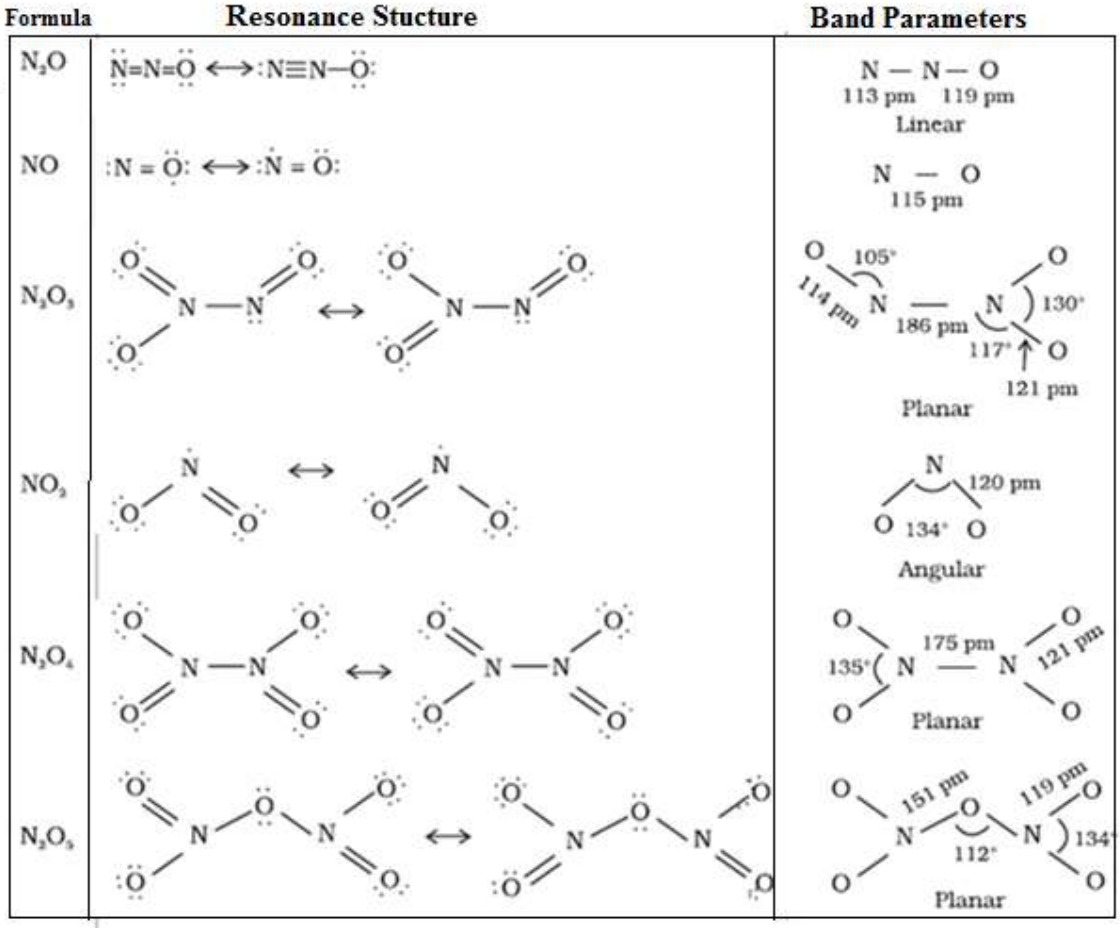
Nitrogen forms oxoacids such as $H_2N_2O_2$ (hyponitrous acid), HNO_2 (nitrous acid) and HNO_3 (nitric acid). Amongst them HNO_3 is the most important.

Preparation

In the laboratory, nitric acid is prepared by heating KNO_3 or $NaNO_3$ and concentrated H_2SO_4 in a glass retort.

$$NaNO_3 + H_2SO_4 \rightarrow NaHSO_4 + HNO_3$$

Lewis dot main resonance structures and bond parameters of oxides are given in the following Table



On a large scale it is prepared mainly by Ostwald's process. This method is based upon catalytic oxidation of NH_3 by atmospheric oxygen.

$$4NH_3(g) + 5O_2(g) \xrightarrow[500\,K,\ 9\,bar]{Pt\,/\,Rh\ gauge\ catalyst} 4NO(g) + 6H_2O(g)$$
(from air)

Nitric oxide thus formed combines with oxygen giving NO_2.

$$2NO(g) + O_2(g) \rightleftharpoons 2NO_2(g)$$

Nitrogen dioxide so formed, dissolves in water to give HNO_3.

$$3NO_2(g) + H_2O(l) \rightarrow 2HNO_3(aq) + NO(g)$$

NO thus formed is recycled and the aqueous HNO_3 can be concentrated by distillation upto ~ 68% by mass. Further concentration to 98% can be achieved by dehydration with concentrated H_2SO_4.

Properties

It is a colourless liquid (f.p. 231.4 K and b.p. 355.6 K). Laboratory grade nitric acid contains ~ 68% of the HNO_3 by mass and has a specific gravity of 1.504. In the gaseous state, HNO_3 exists as a planar molecule with the structure as shown. In aqueous solution, nitric acid behaves as a strong acid giving hydronium and nitrate ions.

$$HNO_3(aq) + H_2O(l) \rightarrow H_3O^+(aq) + NO_3^-(aq)$$

Concentrated nitric acid is a strong oxidising agent and attacks most metals except noble metals such as gold and platinum. The products of oxidation depend upon the concentration of the acid, temperature and the nature of the material undergoing oxidation.

$$3Cu + 8\,HNO_3(\text{dilute}) \rightarrow 3Cu(NO_3)_2 + 2NO + 4H_2O$$

$$Cu + 4HNO_3(\text{conc.}) \rightarrow Cu(NO_3)_2 + 2NO_2 + 2H_2O$$

Zinc reacts with dilute nitric acid to give N2O and with concentrated acid to give NO_2.

$$4Zn + 10HNO_3(\text{dilute}) \rightarrow 4\,Zn\,(NO_3)_2 + 5H_2O + N_2O$$

$$Zn + 4HNO_3(\text{conc.}) \rightarrow Zn\,(NO_3)_2 + 2H_2O + 2NO_2$$

Some metals (e.g., Cr, Al) do not dissolve in concentrated nitric acid because of the formation of a passive film of oxide on the surface. Concentrated nitric acid also oxidizes non–metals and their compounds. Iodine is oxidised to iodic acid, carbon to carbon dioxide, sulphur to H_2SO_4, and phosphorus to phosphoric acid.

$$I_2 + 10HNO_3 \rightarrow 2HIO_3 + 10\,NO_2 + 4H_2O$$

$$C + 4HNO_3 \rightarrow CO_2 + 2H_2O + 4NO_2$$

$$S_8 + 48HNO_3(\text{conc.}) \rightarrow 8H_2SO_4 + 48NO_2 + 16H_2O$$

$$P_4 + 20HNO_3(\text{conc.}) \rightarrow 4H_3PO_4 + 20\,NO_2 + 4H_2O$$

Brown Ring Test

The familiar brown ring test for nitrates depends on the ability of Fe2+ to reduce nitrates to nitric oxide, which reacts with Fe2+ to form a brown coloured complex. The test is usually carried out by adding dilute ferrous sulphate solution to an aqueous solution containing nitrate ion, and then carefully adding concentrated sulphuric acid along the sides of the test tube. A brown ring at the interface between the solution and sulphuric acid layers indicate the presence of nitrate ion in solution.

$$NO_3^- + 3Fe^{2+} + 4H^+ \rightarrow NO + 3Fe^{3+} + 2H_2O$$

$$[Fe(H_2O)_6]^{2+} + NO \rightarrow [Fe\,(H_2O)_5\,(NO)]^{2+} + H_2O$$

(brown)

6.6 Phosphorus Allotropic Forms

Phosphorus is found in many allotropic forms, the important ones being white, red and black. **White phosphorus** is a translucent white waxy solid. It is poisonous, insoluble in water but soluble in carbon disulphide and glows in dark (chemiluminescence). It dissolves in boiling NaOH solution in an inert atmosphere giving PH_3.

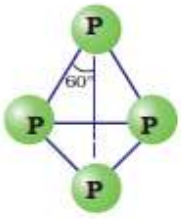
$$P_4 + 3NaOH + 3H_2O \rightarrow PH_3 + 3NaH_2PO_2$$

(sodium hypophosphite)

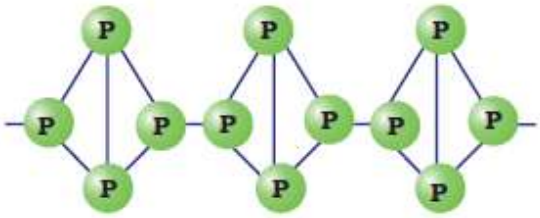
White phosphorus is less stable and therefore, more reactive than the other solid phases under normal conditions because of angular strain in the P_4 molecule where the angles are only 60°. It readily catches fire in air to give dense white fumes of P_4O_{10}.

$$P_4 + 5O_2 \rightarrow P_4O_{10}$$

It consists of discrete tetrahedral P4 molecule as shown in Fig. **Red phosphorus** is obtained by heating white phosphorus at 573K in an inert atmosphere for several days. When red phosphorus is heated under high pressure, a series of phases of black phosphorus are formed. Red phosphorus possesses iron grey lustre. It is odourless, nonpoisonous and insoluble in water as well as in carbon disulphide. Chemically, red phosphorus is much less reactive than white phosphorus. It does not glow in the dark.



White phosphorus



Red phosphorus

Black phosphorus has two forms α-black phosphorus and β-black phosphorus. α-Black phosphorus is formed when red phosphorus is heated in a sealed tube at 803K. It can be sublimed in air and has opaque monoclinic or rhombohedral crystals. It does not oxidise in air. β-Black phosphorus is prepared by heating white phosphorus at 473 K under high pressure. It does not burn in air upto 673 K.

6.7 Phosphine

Preparation

Phosphine is prepared by the reaction of calcium phosphide with water or dilute HCl.

$$Ca_3P_2 + 6H_2O \rightarrow 3Ca(OH)_2 + 2PH_3$$

$$Ca_3P_2 + 6HCl \rightarrow 3CaCl_2 + 2PH_3$$

In the laboratory, it is prepared by heating white phosphorus with concentrated NaOH solution in an inert atmosphere of CO_2.

$$P_4 + 3NaOH + 3H_2O \rightarrow PH_3 + 3NaH_2PO_2$$

(sodium hypophosphite)

Properties

It is a colourless gas with rotten fish smell and is highly poisonous. It explodes in contact with traces of oxidising agents like HNO_3, Cl_2 and Br_2 vapours. It is slightly soluble in water. The solution of PH_3 in water decomposes in presence of light giving red phosphorus and H_2. When absorbed in copper sulphate or mercuric chloride solution, the corresponding phosphides are obtained.

$$3CuSO_4 + 2PH_3 \rightarrow Cu_3P_2 + 3H_2SO_4$$
$$3HgCl_2 + 2PH_3 \rightarrow Hg_3P_2 + 6HCl$$

Phosphine is weakly basic and like ammonia, gives phosphonium compounds with acids e.g.,

$$PH_3 + HBr \rightarrow PH_4Br$$

6.8 Phosphorus Halides

Phosphorus forms two types of halides, PX_3 (X = F, Cl, Br, I) and PX_5 (X = F, Cl, Br).

6.8.1 Phosphorus Trichloride

Preparation

It is obtained by passing dry chlorine over heated white phosphorus.

$$P_4 + 6Cl_2 \rightarrow 4PCl_3$$

Properties

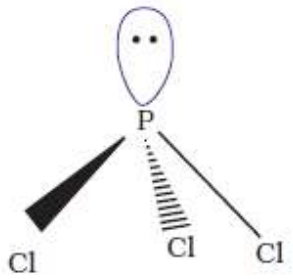
It is a colourless oily liquid and hydrolyses in the presence of moisture.

$$PCl_3 + 3H_2O \rightarrow H_3PO_3 + 3HCl$$

It reacts with organic compounds containing –OH group such as CH_3COOH, C_2H_5OH.

$$3CH_3COOH + PCl_3 \rightarrow 3CH_3COCl + H_3PO_3$$
$$3C_2H_5OH + PCl_3 \rightarrow 3C_2H_5Cl + H_3PO_3$$

Structure of PCl_3



It has a pyramidal shape as shown, in which phosphorus is *sp*3 hybridised.

6.8.2 Phosphorus Pentachloride

Preparation

Phosphorus pentachloride is prepared by the reaction of white phosphorus with excess of dry chlorine.

$$P_4 + 10Cl_2 \rightarrow 4PCl_5$$

It can also be prepared by the action of SO_2Cl_2 on phosphorus.

$$P_4 + 10SO_2Cl_2 \rightarrow 4PCl_5 + 10SO_2$$

Properties

PCl_5 is a yellowish white powder and in moist air, it hydrolyses to $POCl_3$ and finally gets converted to phosphoric acid.

$$PCl_5 + H_2O \rightarrow POCl_3 + 2HCl$$
$$POCl_3 + 3H_2O \rightarrow H_3PO_4 + 3HCl$$

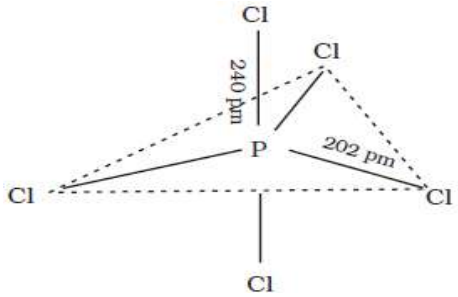
When heated, it sublimes but decomposes on stronger heating.

$$PCl_5 \xrightarrow{\text{Heat}} PCl_3 + Cl_2$$

It reacts with organic compounds containing –OH group converting them to chloro derivatives.

$$C_2H_5OH + PCl_5 \rightarrow C_2H_5Cl + POCl_3 + HCl$$
$$CH_3COOH + PCl_5 \rightarrow CH_3COCl + POCl_3 + HCl$$

Structure of PCl_5



In gaseous and liquid phases, it has a trigonal bipyramidal structure as shown below. The three equatorial P–Cl bonds are equivalent, while the two axial bonds are longer than equatorial bonds. This is due to the fact that the axial bond pairs suffer more repulsion as compared to equatorial bond pairs.

In the solid state it exists as an ionic solid, $[PCl_4]^+[PCl_6]^-$ in which the cation, $[PCl_4]^+$ is tetrahedral and the anion, $[PCl_6]^-$ octahedral.

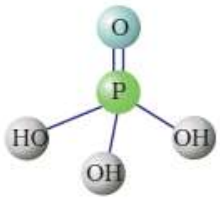
6.9 Oxoacids of Phosphorus

Phosphorus forms a number of oxoacids. The important oxoacids of phosphorus with their formulas, methods of preparation and the presence of some characteristic bonds in their structures are given in the following Table

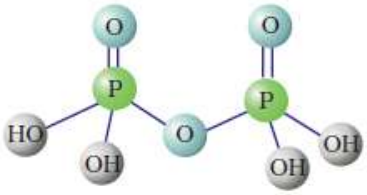
Oxoacids of Phosphorus

Name	Formula	Oxidation state of phosphorus	Characteristic bonds and their number	Preparation
Hypophosphorous (Phosphinic)	H_3PO_2	+1	One P – OH Two P – H One P = O	white P_4 + alkali
Orthophosphorous (Phosphonic)	H_3PO_3	+3	Two P – OH One P – H One P = O	$P_2O_3 + H_2O$
Pyrophosphorous	$H_4P_2O_5$	+3	Two P – OH Two P – H Two P = O	$PCl_3 + H_3PO_3$
Hypophosphoric	$H_4P_2O_6$	+4	Four P – OH Two P = O One P – P	red P_4 + alkali
Orthophosphoric	H_3PO_4	+5	Three P – OH One P = O	$P_4O_{10}+H_2O$
Pyrophosphoric	$H_4P_2O_7$	+5	Four P – OH Two P = O One P – O – P	heat phosphoric acid
Metaphosphoric*	$(HPO_3)_n$	+5	Three P – OH Three P = O Three P – O – P	phosphorus acid + Br_2, heat in a sealed tube

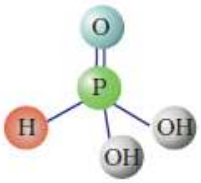
Structures of some important oxoacids of phosphorus



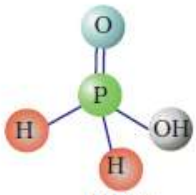
H_3PO_4
Orthophosphoric acid



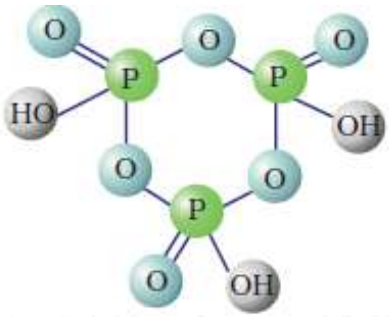
$H_4P_2O_7$
Pyrophosphoric acid



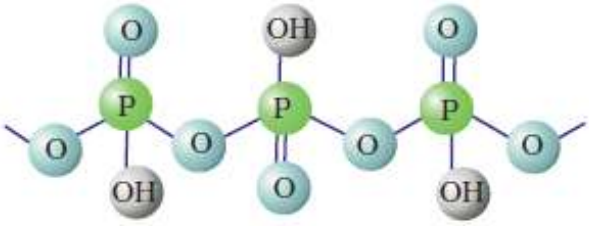
H_3PO_3
Orthophosphorous acid



H_3PO_2
Hypophosphorous acid



Cyclotrimetaphosphoric acid, $(HPO_3)_3$



Polymetaphosphoric acid, $(HPO_3)_n$

GROUP 16 ELEMENTS

Oxygen, sulphur, selenium, tellurium and polonium constitute Group 16 of the periodic table. This is sometimes known as group of *chalcogens*. The name is derived from the Greek word for brass and points to the association of sulphur and its congeners with copper. Most copper minerals contain either oxygen or sulphur and frequently the other members of the group.

6.10 Occurrence

Oxygen is the most abundant of all the elements on earth. Oxygen forms about 46.6% by mass of earth's crust. Dry air contains 20.946% oxygen by volume.

However, the abundance of sulphur in the earth's crust is only 0.03-0.1%. Combined sulphur exists primarily as sulphates such as *gypsum* $CaSO_4.2H_2O$, *epsom salt* $MgSO_4.7H_2O$, *baryte* $BaSO_4$ and sulphides such as *galena* PbS, *zinc blende* ZnS, *copper pyrites* $CuFeS_2$. Traces of sulphur occur as hydrogen sulphide in volcanoes. Organic materials such as eggs, proteins, garlic, onion, mustard, hair and wool contain sulphur.

Selenium and tellurium are also found as metal selenides and tellurides in sulphide ores. Polonium occurs in nature as a decay product of thorium and uranium minerals.

Electronic Configuration

The elements of Group16 have six electrons in the outermost shell and have ns^2np^4 general electronic configuration.

Atomic and Ionic Radii

Due to increase in the number of shells, atomic and ionic radii increase from top to bottom in the group. The size of oxygen atom is, however, exceptionally small.

Ionisation Enthalpy

Ionisation enthalpy decreases down the group. It is due to increase in size. However, the elements of this group have lower ionisation enthalpy values compared to those of Group15 in the corresponding periods. This is due to the fact that Group 15 elements have extra stable halffilled *p* orbitals electronic configurations.

Electron Gain Enthalpy

Because of the compact nature of oxygen atom, it has less negative electron gain enthalpy than sulphur. However, from sulphur onwards the value again becomes less negative upto polonium.

Electronegativity

Next to fluorine, oxygen has the highest electronegativity value amongst the elements. Within the group, electronegativity decreases with an increase in atomic number. This implies that the metallic character increases from oxygen to polonium.

Physical Properties

Oxygen and sulphur are non-metals, selenium and tellurium metalloids, whereas polonium is a metal. Polonium is radioactive and is short lived (Half-life 13.8 days). All these elements exhibit allotropy. The melting and boiling points increase with an increase in atomic number down the group. The large difference between the melting and boiling points of oxygen and sulphur may be explained on the basis of their atomicity; oxygen exists as diatomic molecule (O_2) whereas sulphur exists as polyatomic molecule (S_8).

Chemical Properties

Oxidation states and trends in chemical reactivity

The elements of Group 16 exhibit a number of oxidation states. The stability of -2 oxidation state decreases down the group. Polonium hardly shows –2 oxidation state. Since electronegativity of oxygen is very high, it shows only negative oxidation state as –2 except in the case of OF_2 where its oxidation state is + 2. Other elements of the group exhibit + 2, + 4, + 6 oxidation states but + 4 and + 6 are more common. Sulphur, selenium and tellurium usually show + 4 oxidation state in their compounds with oxygen and + 6 with fluorine. The stability of + 6 oxidation state decreases down the group and stability of + 4 oxidation state increase (inert pair effect). Bonding in +4 and +6 oxidation states are primarily covalent.

Anomalous behaviour of oxygen

The anomalous behaviour of oxygen, like other members of *p*-block present in second period is due to its small size and high electronegativity. One typical example of effects of small size and high electronegativity is the presence of strong hydrogen bonding in H_2O which is not found in H_2S.

The absence of *d* orbitals in oxygen limits its covalency to four and in practice, rarely exceeds two. On the other hand, in case of other elements of the group, the valence shells can be expanded and covalence exceeds four.

(i) Reactivity with hydrogen:

All the elements of Group 16 form hydrides of the type H_2E (E = S, Se, Te, Po). Some properties of hydrides are given in the following Table. Their acidic character increases from H_2O to H_2Te. The increase in acidic character can be explained in terms of decrease in bond (H–E) dissociation enthalpy down the group. Owing to the decrease in bond (H–E) dissociation enthalpy down the group, the thermal stability of hydrides also decreases from H_2O to H_2Po. All the hydrides except water possess reducing property and this character increases from H_2S to H_2Te.

Properties of Hydrides of Group 16 Elements

Property	H_2O	H_2S	H_2Se	H_2Te
m.p/K	273	188	208	222
b.p/K	373	213	232	269
H–E distance/pm	96	134	146	169
HEH angle (°)	104	92	91	90
$\Delta_f H$/kJ mol^{-1}	–286	–20	73	100
$\Delta_{diss} H$ *(H–E)*/kJ mol^{-1}	463	347	276	238
Dissociation constant[a]	1.8×10^{-16}	1.3×10^{-7}	1.3×10^{-4}	2.3×10^{-3}

[a] *Aqueous solution, 298 K*

(ii) Reactivity with oxygen:

All these elements form oxides of the EO_2 and EO_3 types where E = S, Se, Te or Po. Ozone (O_3) and sulphur dioxide (SO_2) are gases while selenium dioxide (SeO_2) is solid. Reducing property of dioxide decreases from SO_2 to TeO_2; SO_2 is reducing while TeO_2 is an oxidising agent. Besides EO_2 type, sulphur, selenium and tellurium also form EO_3 type oxides (SO_3, SeO_3, TeO_3). Both types of oxides are acidic in nature.

(iii) Reactivity towards the halogens:

Elements of Group 16 form a large number of halides of the type, EX_6, EX_4 and EX_2 where E is an element of the group and X is a halogen. The stability of the halides decreases in the order F– > Cl– > Br– > I–. Amongst hexahalides, hexafluorides are the only stable halides. All hexafluorides are gaseous in nature. They have octahedral structure. Sulphur hexafluoride, SF_6 is exceptionally stable for steric reasons.

Amongst tetrafluorides, SF_4 is a gas, SeF_4 a liquid and TeF_4 a solid. These fluorides have sp^3d hybridisation and thus, have trigonal bipyramidal structures in which one of the equatorial positions is occupied by a lone pair of electrons. This geometry is also regarded as *see-saw* geometry.

All elements except selenium form dichlorides and dibromides. These dihalides are formed by *sp*3 hybridisation and thus, have tetrahedral structure. The well known monohalides are dimeric in nature. Examples are S_2F_2, S_2Cl_2, S_2Br_2, Se_2Cl_2 and Se_2Br_2. These dimeric halides undergo disproportionation as given below:

$$2Se_2Cl_2 \rightarrow SeCl_4 + 3Se$$

6.11 Simple Oxides

A binary compound of oxygen with another element is called oxide. As already stated, oxygen reacts with most of the elements of the periodic table to form oxides. In many cases one element forms two or more oxides. The oxides vary widely in their nature and properties.

Oxides can be simple (e.g., MgO, Al_2O_3) or mixed (Pb_3O_4, $_{Fe3O4}$). Simple oxides can be classified on the basis of their acidic, basic or amphoteric character. An oxide that combines with water to give an acid is termed acidic oxide (e.g., SO_2, Cl_2O_7, CO_2, N_2O_5). For example, SO_2 combines with water to give H_2SO_3, an acid.

$$SO_2 + H_2O \rightarrow H_2SO_3$$

In general, metallic oxides are basic. Some metallic oxides exhibit a dual behaviour. They show characteristics of both acidic as well as basic oxides. Such oxides are known as amphoteric oxides. They react with acids as well as alkalies. There are some oxides which are neither acidic nor basic. Such oxides are known as neutral oxides. Examples of neutral oxides are CO, NO and N_2O. For example, Al_2O_3 reacts with acids as well as alkalies.

$$Al_2O_3(s) + 6HCl(aq) + 9H_2O(l) \rightarrow 2[Al(H_2O)_6]^{3+}(aq) + 6Cl^-(aq)$$
$$Al_2O_3(s) + 6NaOH(aq) + 3H_2O(l) \rightarrow 2Na_3[Al(OH)_6](aq)$$

6.12 Ozone

Ozone is an allotropic form of oxygen. It is too reactive to remain for long in the atmosphere at sea level. At a height of about 20 kilometres, it is formed from atmospheric oxygen in the presence of sunlight. This ozone layer protects the earth's surface from an excessive concentration of ultraviolet (UV) radiations.

Preparation

When a slow dry stream of oxygen is passed through a silent electrical discharge, conversion of oxygen to ozone (10%) occurs. The product is known as ozonised oxygen.

$$3O_2 \rightarrow 2O_3 \quad \Delta H^\ominus (298\ K) = +142\ kJ\ mol^{-1}$$

Since the formation of ozone from oxygen is an endothermic process, it is necessary to use a silent electrical discharge in its preparation to prevent its decomposition.

Properties

Pure ozone is a pale blue gas, dark blue liquid and violet-black solid. Ozone has a characteristic smell and in small concentrations it is harmless. However, if the concentration rises above about 100 parts per million, breathing becomes uncomfortable resulting in headache and nausea.

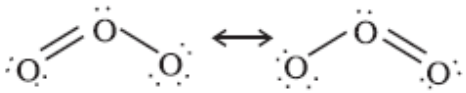
Due to the ease with which it liberates atoms of nascent oxygen ($O_3 \rightarrow O_2 + O$), it acts as a powerful oxidising agent. For e.g., it oxidizes lead sulphide to lead sulphate and iodide ions to iodine.

$$PbS(s) + 4O_3(g) \rightarrow PbSO_4(s) + 4O_2(g)$$
$$2I^-(aq) + H_2O(l) + O_3(g) \rightarrow 2OH^-(aq) + I_2(s) + O_2(g)$$

When ozone reacts with an excess of potassium iodide solution buffered with a borate buffer (pH 9.2), iodine is liberated which can be titrated against a standard solution of sodium thiosulphate. This is a quantitative method for estimating O_3 gas.

Structure

The two oxygen-oxygen bond lengths in the ozone molecule are identical (128 pm) and the molecule is angular as expected with a bond angle of about 117o. It is a resonance hybrid of two main forms:

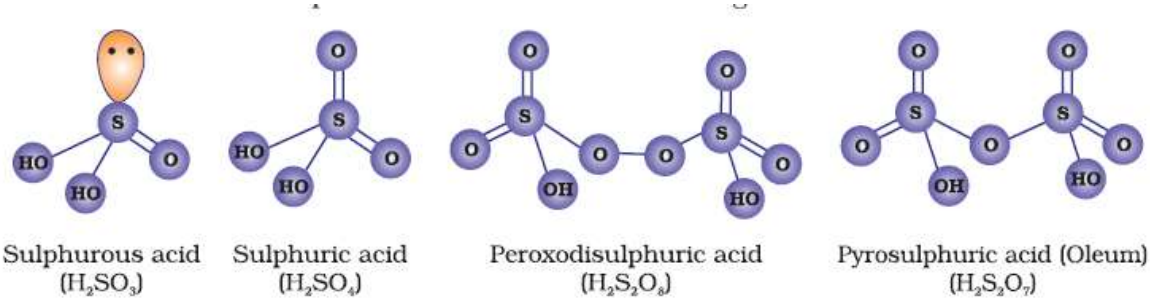


Uses:

It is used as a germicide, disinfectant and for sterilising water. It is also used for bleaching oils, ivory, flour, starch, etc. It acts as an oxidising agent in the manufacture of potassium permanganate.

6.13 Oxoacids of Sulphur

Sulphur forms a number of oxoacids such as H_2SO_3, $H_2S_2O_3$, $H_2S_2O_4$, $H_2S_2O_5$, H_2SxO_6 (x = 2 to 5), H_2SO_4, $H_2S_2O_7$, H_2SO_5, $H_2S_2O_8$. Some of these acids are unstable and cannot be isolated. They are known in aqueous solution or in the form of their salts. Structures of some important oxoacids are shown in the following Figure



Structures of some important oxoacids of sulphur

GROUP 17 ELEMENTS

Fluorine, chlorine, bromine, iodine and astatine are members of Group 17. These are collectively known as the **halogens** (Greek *halo* means salt and *genes* means born i.e., salt producers). The halogens are highly reactive non-metallic elements. Like Groups 1 and 2, the elements of Group 17 show great similarity amongst themselves. That much similarity is not found in the elements of other groups of the periodic table. Also, there is a regular gradation in their physical and chemical properties. Astatine is a radioactive element.

6.14 Occurrence

Fluorine and chlorine are fairly abundant while bromine and iodine less so. Fluorine is present mainly as insoluble fluorides (fluorspar CaF_2, cryolite Na_3AlF_6 and fluoroapatite $3Ca_3(PO_4)_2.CaF_2$) and small quantities are present in soil, river water plants and bones and teeth of animals. Sea water contains chlorides, bromides and iodides of sodium, potassium, magnesium and calcium, but is mainly sodium chloride solution (2.5% by mass). The deposits of dried up seas contain these compounds, e.g., sodium chloride and carnallite, $KCl.MgCl_2.6H_2O$. Certain forms of marine life contain iodine in their systems; various seaweeds, for example, contain upto 0.5% of iodine and Chile saltpetre contains upto 0.2% of sodium iodate

Electronic configuration

All these elements have seven electrons in their outermost shell (ns^2np^5) which is one electron short of the next noble gas.

Atomic and ionic radii

The halogens have the smallest atomic radii in their respective periods due to maximum effective nuclear charge. The atomic radius of fluorine like the other elements of second period is extremely small. Atomic and ionic radii increase from fluorine to iodine due to increasing number of quantum shells.

Ionisation enthalpy

They have little tendency to lose electron. Thus they have very high ionisation enthalpy. Due to increase in atomic size, ionisation enthalpy decreases down the group.

Electron gain enthalpy

Halogens have maximum negative electron gain enthalpy in the corresponding periods. This is due to the fact that the atoms of these elements have only one electron less than stable noble gas configurations. Electron gain enthalpy of the elements of the group becomes less negative down the group. However, the negative electron gain enthalpy of fluorine is less than that of chlorine. It is due to small size of fluorine atom. As a result, there are strong interelectronic repulsions in the relatively small 2*p* orbitals of fluorine and thus, the incoming electron does not experience much attraction.

Electronegativity

They have very high electronegativity. The electronegativity decreases down the group. Fluorine is the most electronegative element in the periodic table.

6.15 Physical and chemical properties

Physical properties

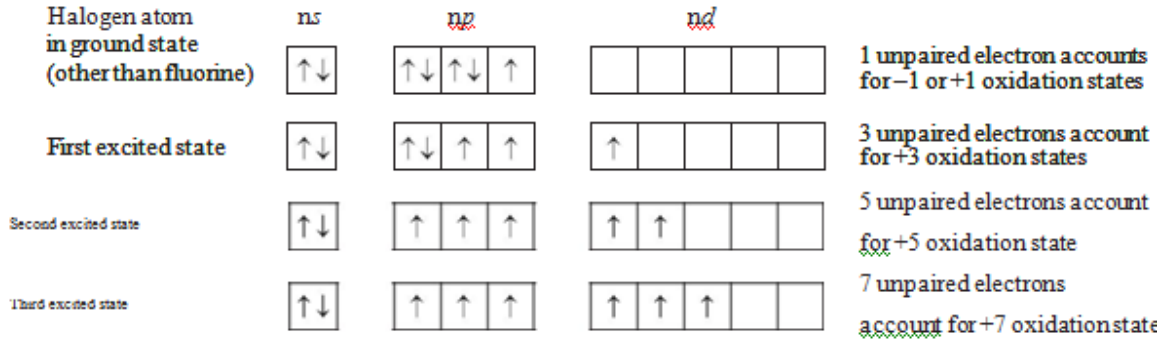
Halogens display smooth variations in their physical properties. Fluorine and chlorine are gases, bromine is a liquid and iodine is a solid. Their melting and boiling points steadily increase with atomic number. All halogens are coloured. This is due to absorption of radiations in visible region which results in the excitation of outer electrons to higher energy level. By absorbing different quanta of radiation, they display different colours. For example, F_2, has yellow, Cl_2 ,

greenish yellow, Br_2, red and I_2, violet colour. Fluorine and chlorine react with water. Bromine and iodine are only sparingly soluble in water but are soluble in various organic solvents such as chloroform, carbon tetrachloride, carbon disulphide and hydrocarbons to give coloured solutions. One curious anomaly we notice is the smaller enthalpy of dissociation of F_2 compared to that of Cl_2 whereas X-X bond dissociation enthalpies from chlorine onwards show the expected trend: Cl – Cl > Br – Br > I – I. A reason for this anomaly is the relatively large electron-electron repulsion among the lone pairs in F_2 molecule where they are much closer to each other than in case of Cl_2.

Chemical properties

Oxidation states and trends in chemical reactivity

All the halogens exhibit –1 oxidation state. However, chlorine, bromine and iodine exhibit + 1, + 3, + 5 and + 7 oxidation states also as explained below



The higher oxidation states of chlorine, bromine and iodine are realised mainly when the halogens are in combination with the small and highly electronegative fluorine and oxygen atoms. e.g., in interhalogens, oxides and oxoacids. The oxidation states of +4 and +6 occur in the oxides and oxoacids of chlorine and bromine. The fluorine atom has no *d* orbitals in its valence shell and therefore cannot expand its octet. Being the most electronegative, it exhibits only –1 oxidation state.

All the halogens are highly reactive. They react with metals and non-metals to form halides. The reactivity of the halogens decreases down the group.

The ready acceptance of an electron is the reason for the strong oxidising nature of halogens. F_2 is the strongest oxidising halogen and it oxidises other halide ions in solution or even in the solid phase. In general, a halogen oxidises halide ions of higher atomic number.

$$F_2 + 2X^- \rightarrow 2F^- + X_2 \quad (X = Cl, Br \text{ or } I)$$
$$Cl_2 + 2X^- \rightarrow 2Cl^- + X_2 \quad (X = Br \text{ or } I)$$
$$Br_2 + 2I^- \rightarrow 2Br^- + I_2$$

The relative oxidising power of halogens can further be illustrated by their reactions with water. Fluorine oxidises water to oxygen whereas chlorine and bromine react with water to form corresponding hydrohalic and hypohalous acids. The reaction of iodine with water is non-spontaneous. In fact, I^- can be oxidised by oxygen in acidic medium; just the reverse of the reaction observed with fluorine.

$$2F_2\,(g) + 2H_2O\,(l) \rightarrow 4H^+\,(aq) + 4F^-\,(aq) + O_2\,(g)$$

$$X_2\,(g) + H_2O\,(l) \rightarrow HX\,(aq) + HOX\,(aq)$$
(where X = Cl or Br)

$$4I^-\,(aq) + 4H^+\,(aq) + O_2\,(g) \rightarrow 2I_2\,(s) + 2H_2O\,(l)$$

Anomalous behaviour of fluorine:

Like other elements of *p*-block present in second period of the periodic table, fluorine is anomalous in many properties. For example, ionisation enthalpy, electronegativity, enthalpy of bond dissociation and electrode potentials are all higher for fluorine than expected from the trends set by other halogens. Also, ionic and covalent radii, m.p. and b.p. and electron gain enthalpy are quite lower than expected. The anomalous behaviour of fluorine is due to its small size, highest electronegativity, low F-F bond dissociation enthalpy, and non availability of *d* orbitals in valence shell.

Most of the reactions of fluorine are exothermic (due to the small and strong bond formed by it with other elements). It forms only one oxoacid while other halogens form a number of oxoacids. Hydrogen fluoride is a liquid (b.p. 293 K) due to strong hydrogen bonding. Other hydrogen halides are gases.

Reactivity towards hydrogen

They all react with hydrogen to give hydrogen halides but affinity for hydrogen decreases from fluorine to iodine. They dissolve in water to form hydrohalic acids. The acidic strength of these acids varies in the order: HF < HCl < HBr < HI. The stability of these halides decreases down the group due to decrease in bond (H–X) dissociation enthalpy in the order:

H–F > H–Cl > H–Br > H–I.

Reactivity towards oxygen

Halogens form many oxides with oxygen but most of them are unstable. Fluorine forms two oxides OF_2 and O_2F_2. However, only OF_2 is thermally stable at 298 K. These oxides are essentially oxygen fluorides because of the higher electronegativity of fluorine than oxygen. Both are strong fluorinating agents. O_2F_2 oxidises plutonium to PuF_6 and the reaction is used in removing plutonium as PuF_6 from spent nuclear fuel.

Chlorine, bromine and iodine form oxides in which the oxidation states of these halogens range from +1 to +7. A combination of kinetic and thermodynamic factors lead to the generally decreasing order of stability of oxides formed by halogens, I > Cl > Br. The higher oxides of halogens tend to be more stable than the lower ones.

Chlorine oxides, Cl_2O, ClO_2, Cl_2O_6 and Cl_2O_7 are highly reactive oxidising agents and tend to explode. ClO_2 is used as a bleaching agent for paper pulp and textiles and in water treatment.

The bromine oxides, Br_2O, BrO_2, BrO_3 are the least stable halogen oxides (middle row anomally) and exist only at low temperatures. They are very powerful oxidising agents.

The iodine oxides, I_2O_4 , I_2O_5, I_2O_7 are insoluble solids and decompose on heating. I_2O_5 is a very good oxidising agent and is used in the estimation of carbon monoxide.

Reactivity towards metals

Halogens react with metals to form metal halides. For e.g., bromine reacts with magnesium to give magnesium bromide.

$$Mg\,(s) + Br_2\,(l) \rightarrow MgBr_2\,(s)$$

The ionic character of the halides decreases in the order MF > MCl > MBr > MI where M is a monovalent metal. If a metal exhibits more than one oxidation state, the halides in higher oxidation state will be more covalent than the one in lower oxidation state. For e.g., $SnCl_4$, $PbCl_4$, $SbCl_5$ and UF_6 are more covalent than $SnCl_2$, $PbCl_2$, $SbCl_3$ and UF_4 respectively.

Reactivity of halogens towards other halogens

Halogens combine amongst themselves to form a number of compounds known as interhalogens of the types XX', XX_3', XX_5' and XX_7' where X is a larger size halogen and X' is smaller size halogen.

6.16 Chlorine

Chlorine was discovered in 1774 by Scheele by the action of HCl on MnO_2. In 1810 Davy established its elementary nature and suggested the name chlorine on account of its colour (Greek, *chloros* = greenish yellow).

Preparation

It can be prepared by any one of the following methods:

By heating manganese dioxide with concentrated hydrochloric acid. MnO_2

$$4HCl \rightarrow MnCl_2 + Cl_2 + 2H_2O$$

However, a mixture of common salt and concentrated H_2SO_4 is used in place of HCl

$$4NaCl + MnO_2 + 4H_2SO_4 \rightarrow MnCl_2 + 4NaHSO_4 + 2H_2O + Cl_2$$

By the action of HCl on potassium permanganate. $2KMnO_4$

$$16HCl \rightarrow 2KCl + 2MnCl_2 + 8H_2O + 5Cl_2$$

Manufacture of chlorine

(i) Deacon's process: By oxidation of hydrogen chloride gas by atmospheric oxygen in the presence of $CuCl_2$ (catalyst) at 723 K.

$$4HCl + O_2 \xrightarrow{CuCl_2} 2Cl_2 + 2H_2O$$

(ii) Electrolytic process: Chlorine is obtained by the electrolysis of brine (concentrated NaCl solution). Chlorine is liberated at anode. It is also obtained as a by–product in many chemical industries.

Properties

It is a greenish yellow gas with pungent and suffocating odour. It is about 2-5 times heavier than air. It can be liquefied easily into greenish yellow liquid which boils at 239 K. It is soluble in water.

Chlorine reacts with a number of metals and non-metals to form chlorides

$$2Al + 3Cl_2 \rightarrow 2AlCl_3 ; \qquad P_4 + 6Cl_2 \rightarrow 4PCl_3$$
$$2Na + Cl_2 \rightarrow 2NaCl; \qquad S_8 + 4Cl_2 \rightarrow 4S_2Cl_2$$
$$2Fe + 3Cl_2 \rightarrow 2FeCl_3 ;$$

It has great affinity for hydrogen. It reacts with compounds containing hydrogen to form HCl.

$$H_2 + Cl_2 \rightarrow 2HCl$$
$$H_2S + Cl_2 \rightarrow 2HCl + S$$
$$C_{10}H_{16} + 8Cl_2 \rightarrow 16HCl + 10C$$

With excess ammonia, chlorine gives nitrogen and ammonium chloride whereas with excess chlorine, nitrogen trichloride (explosive) is formed.

$$\underset{\text{(excess)}}{8NH_3} + 3Cl_2 \rightarrow 6NH_4Cl + N_2 \; ; \qquad NH_3 + 3Cl_2 \rightarrow \underset{\text{(excess)}}{NCl_3} + 3HCl$$

With cold and dilute alkalies chlorine produces a mixture of chloride and hypochlorite but with hot and concentrated alkalies it gives chloride and chlorate.

$$\underset{\text{(cold and dilute)}}{2NaOH} + Cl_2 \rightarrow NaCl + NaOCl + H_2O$$

$$\underset{\text{(hot and conc.)}}{6\,NaOH} + 3Cl_2 \rightarrow 5NaCl + NaClO_3 + 3H_2O$$

With dry slaked lime it gives bleaching powder.

$$2Ca(OH)_2 + 2Cl_2 \rightarrow Ca(OCl)_2 + CaCl_2 + 2H_2O$$

Chlorine water on standing loses its yellow colour due to the formation of HCl and HOCl. Hypochlorous acid (HOCl) so formed, gives nascent oxygen which is responsible for oxidising and bleaching properties of chlorine.

It oxidises ferrous to ferric, sulphite to sulphate, sulphur dioxide to sulphuric acid and iodine to iodic acid.

$$2FeSO_4 + H_2SO_4 + Cl_2 \rightarrow Fe_2(SO_4)_3 + 2HCl$$

$$Na_2SO_3 + Cl_2 + H_2O \rightarrow Na_2SO_4 + 2HCl$$

$$SO_2 + 2H_2O + Cl_2 \rightarrow H_2SO_4 + 2HCl$$

$$I_2 + 6H_2O + 5Cl_2 \rightarrow 2HIO_3 + 10HCl$$

It is a powerful bleaching agent; bleaching action is due to oxidation

$$Cl_2 + H_2O \rightarrow 2HCl + O$$

$$\text{Coloured substance} + O \rightarrow \text{Colourless substance}$$

It bleaches vegetable or organic matter in the presence of moisture.Bleaching effect of chlorine is permanent.

Uses

It is used (i) for bleaching woodpulp (required for the manufacture of paper and rayon), bleaching cotton and textiles, (ii) in the extraction of gold and platinum (iii) in the manufacture of dyes, drugs and organic compounds such as CCl_4, $CHCl_3$, DDT, refrigerants, etc. (iv) in sterilising drinking water and (v) preparation of poisonous gases such as phosgene ($COCl_2$), tear gas (CCl_3NO_2), mustard gas ($ClCH_2CH_2SCH_2CH_2Cl$).

6.17 Interhalogen compounds

When two different halogens react with each other, interhalogen compounds are formed. They can be assigned general compositions as XX', XX_3', XX_5' and XX_7' where X is halogen of larger size and X' of smaller size and X is more electropositive than X'. As the ratio between radii of X and X' increases, the number of atoms per molecule also increases. Thus, iodine (VII) fluoride should have maximum number of atoms as the ratio of radii between I and F should be maximum. That is why its formula is IF_7 (having maximum number of atoms).

Preparation

The interhalogen compounds can be prepared by the direct combination or by the action of halogen on lower interhalogen compounds. The product formed depends upon some specific conditions, For e.g.,

$Cl_2 + F_2 \xrightarrow[437\ K]{} 2ClF;$

(equal volume)
$Cl_2 + 3F_2 \xrightarrow[573\ K]{} 2ClF_3;$

(excess)
$I_2 + Cl_2 \rightarrow 2ICl;$
(equimolar)

$I_2 + 3Cl_2 \rightarrow 2ICl_3$

(excess)
$Br_2 + 3F_2 \rightarrow 2BrF_3$

(diluted with water)
$Br_2 + 5F_2 \rightarrow 2BrF_5$
(excess)

Properties

These are all covalent molecules and are diamagnetic in nature. They are volatile solids or liquids except CIF which is a gas at 298 K. Their physical properties are intermediate between those of constituent halogens except that their m.p. and b.p. are a little higher than expected.

Their chemical reactions can be compared with the individual halogens. In general, interhalogen compounds are more reactive than halogens (except fluorine). This is because X–X′ bond in interhalogens is weaker than X–X bond in halogens except F–F bond. All these undergo hydrolysis giving halide ion derived from the smaller halogen and a hypohalite (when XX′), halite (whenXX'_3), halate (when XX'_5) and perhalate (when XX'_7) anion derived from the larger halogen.

$$XX' + H_2O \rightarrow HX' + HOX$$

Their molecular structures are very interesting which can be explained on the basis of VSEPR theory. The XX_3 compounds have the bent 'T' shape, XX_5 compounds square pyramidal and IF_7 has pentagonal bipyramidal structures .

Uses

These compounds can be used as non aqueous solvents. Interhalogen compounds are very useful fluorinating agents. ClF_3 and BrF_3 are used for the production of UF_6 in the enrichment of ^{235}U.

$$U(s) + 3ClF_3(l) \rightarrow UF_6(g) + 3ClF(g)$$

GROUP -18 ELEMENTS

6.18 OCCURRENCE

Group 18 consists of six elements: helium, neon, argon, krypton, xenon and radon. All these are gases and chemically un reactive. They form very few compounds. Because of this they are termed noble gases.

All the noble gases except radon occur in the atmosphere. Their atmospheric abundance in dry air is ~ 1% by volume of which argon is the major constituent. Helium and sometimes neon are found in minerals of radioactive origin e.g., pitchblende, monazite, cleveite. The main commercial source of helium is natural gas. Xenon and radon are the rarest elements of the group. Radon is obtained as a decay product of ^{226}Ra

Electronic configuration

All noble gases have general electronic configuration ns^2np^6 except **Configuration** helium which has$1s.^2$Many of the properties of noble gases including their inactive nature are described to their close shell structures.

Ionisation enthalpy

Due to stable electronic configuration these gases exhibit very high ionisation enthalpy. However, it decreases down the group with increase in atomic size.

Atomic radii

Atomic radii increase down the group with increase in atomic number.

Electron gain enthalpy

Since noble gases have stable electronic configurations, they have no tendency to accept the electron and therefore, have large positive values of electron gain enthalpy

.

6.19 PHYSICAL PROPERTIES

All the noble gases are monoatomic. They are colourless, odourless and tasteless. They are sparingly soluble in water. They have very low melting and boiling points because the only type of interatomic interaction in these elements is weak dispersion forces. Helium has the lowest boiling point (4.2 K) of any known substance. It has an unusual property of diffusing through most commonly used laboratory materials such as rubber, glass or plastics.

Chemical properties

In general, noble gases are least reactive. Their inertness to chemical reactivity is attributed to the following reasons: The noble gases except helium ($1s^2$) have completely filled ns^2np^6 electronic configuration in their valence shell. They have high ionisation enthalpy and more positive electron gain enthalpy.

The reactivity of noble gases has been investigated occasionally, ever since their discoveries, but all attempts to force them to react to form the compounds, were unsuccessful for quite a few years. In March 1962, Neil Bartlett, then at the University of British Columbia, observed the reaction of a noble gas. First, he prepared a red compound which is formulated as $O_2^+PtF_6^-$. He, then realised that the first ionisation enthalpy of molecular oxygen (1175 $kJmol^{-1}$) was almost identical with that of xenon (1170 kJ mol^{-1}). He made efforts to prepare same type of compound with Xe and was successful in preparing another red colour compound $Xe^+PtF_6^-$ by mixing PtF_6 and xenon. After this discovery, a number of xenon compounds mainly with most electronegative elements like fluorine and oxygen, have been synthesised. The compounds of krypton are fewer. Only the difluoride (KrF_2) has been studied in detail. Compounds of radon have not been isolated but only identified (e.g., RnF_2) by radiotracer technique. No true compounds of Ar, Ne or He are yet known.

6.20 Xenon-fluorine compounds

Xenon forms three binary fluorides, XeF_2, XeF_4 and XeF_6 by the direct reaction of elements under appropriate experimental conditions.

$$Xe\,(g) + F_2\,(g) \xrightarrow{673\ K,\ 1\ bar} XeF_2$$
(xenon in excess)

$$Xe\,(g) + 2F_2\,(g) \xrightarrow{873\ K,\ 7\ bar} XeF_4$$
(1:5 ratio)

$$Xe\,(g) + 3F_2\,(g) \xrightarrow{573\ K,\ 60-70\ bar} XeF_6$$
(1:20 ratio)

XeF_6 can also be prepared by the interaction of XeF_4 and O_2F_2 at 143K.

$$XeF_4 + O_2F_2 \rightarrow XeF_6 + O_2$$

XeF_2, XeF_4 and XeF_6 are colourless crystalline solids and sublime readily at 298 K. They are powerful fluorinating agents. They are readily hydrolysed even by traces of water. For example, XeF_2 is hydrolysed to give Xe, HF and O_2.

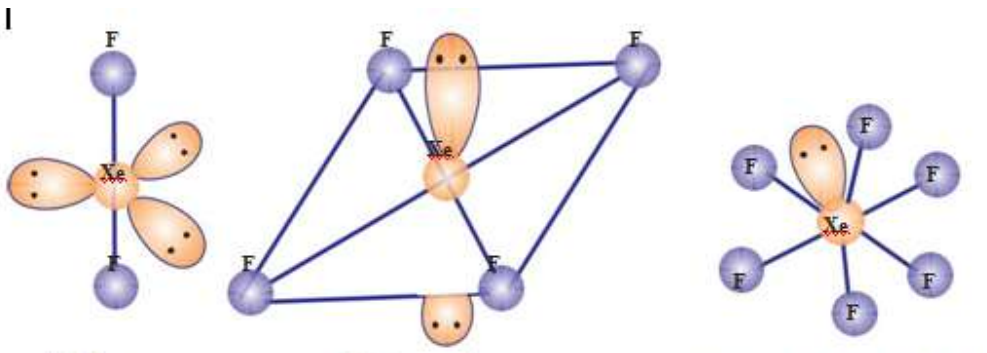
$$2XeF_2\,(s) + 2H_2O(l) \rightarrow 2Xe\,(g) + 4\,HF(aq) + O_2(g)$$

The structures of the three xenon fluorides can be deduced from VSEPR and XeF_2 and XeF_4 have linear and square planar structures respectively. XeF_6 has seven electron pairs (6 bonding pairs and one lone pair) and would, thus, have a distorted octahedral structure as found experimentally in the gas phase

.

Xenon fluorides react with fluoride ion acceptors to form cationic species and fluoride ion donors to form fluoroanions.

$$XeF_2 + PF_5 \rightarrow [XeF]^+[PF_6]^- ; XeF_4 + SbF_5 \rightarrow [XeF_3]^+[SbF_6]^- XeF_6 +$$
$$MF \rightarrow M^+[XeF_7]^- (M = Na, K, Rb \text{ or } Cs)$$



(a) Linear (b) Square planar (c) Distorted octahedral (d)

Xenon-oxygen compounds

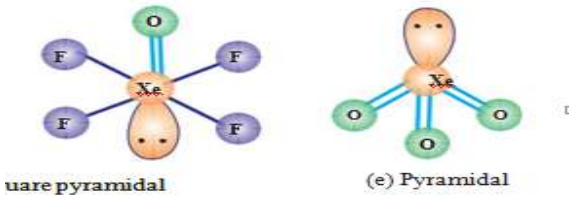
Hydrolysis of XeF_4 and XeF_6 with water gives $Xe0_3$

$$6XeF_4 + 12\,H_2O \rightarrow 4Xe + 2Xe0_3 + 24\,HF + 3\,O_2$$
$$XeF_6 + 3\,H_2O \rightarrow XeO_3 + 6\,HF$$

Partial hydrolysis of XeF_6 gives oxyfluorides, $XeOF_4$ and XeO_2F_2.

$$XeF_6 + H_2O \rightarrow XeOF_4 + 2\,HF$$
$$XeF_6 + 2\,H_2O \rightarrow XeO_2F_2 + 4HF$$

XeO_3 is a colourless explosive solid and has a pyramidal molecular structure.$XeOF_4$ is a colourless volatile liquid and has a square pyramidal molecular structure.



Uses:

Helium is a non-inflammable and light gas. Hence, it is used in filling balloons for meteorological observations. It is also used in gas-cooled nuclear reactors. Liquid helium (b.p. 4.2 K) finds use as cryogenic agent for carrying out various experiments at low temperatures. It is used to produce and sustain powerful superconducting magnets which form an essential part of modern NMR spectrometers and Magnetic Resonance Imaging (MRI) systems for clinical diagnosis. It is used as a diluent for oxygen in modern diving apparatus because of its very low solubility in blood. Neon is used in discharge tubes and fluorescent bulbs for advertisement display purposes. Neon bulbs are used in botanical gardens and in green houses Argon is used mainly to provide an inert atmosphere in high temperature metallurgical processes (arc welding of metals or alloys) and for filling electric bulbs. It is also used in the laboratory for handling substances that are air-sensitive. There are no significant uses of Xenon and Krypton. They are used in light bulbs designed for special purposes

Summary

Groups 13 to 18 of the periodic table consist of ***p*-block elements** with their valence shell electronic configuration ns^2np^{1-6}. Groups 13 and 14 were dealt with in Class XI. In this Unit remaining groups of the *p*-block have been discussed.

Group 15 consists of five elements namely, N, P, As, Sb and Bi which have general electronic configuration ns^2np^3. Nitrogen differs from other elements of this group due to small size, formation of ***pπ–pπ* multiple bonds** with itself and with highly electronegative atom like O or C and **non-availability of *d* orbitals** to expand its valence shell. Elements of group 15 show gradation in properties. They react with oxygen, hydrogen and halogens. They exhibit two important oxidation states, + 3 and + 5 but +3 oxidation is favoured by heavier elements due to 'inert pair effect'.

Dinitrogen can be prepared in laboratory as well as on industrial scale. It forms oxides in various oxidation states as N_2O, NO, N_2O_3, NO_2, N_2O_4 and N_2O_5. These oxides have **resonating structures** and have multiple bonds. Ammonia can be prepared on large scale by **Haber's process**. HNO_3 is an important industrial chemical. It is a strong monobasic acid and is a powerful oxidising agent. Metals and non-metals react with HNO_3 under different conditions to give NO or NO_2.

Phosphorus exists as P_4 in elemental form. It exists in several **allotropic forms**. It forms hydride, PH_3 which is a highly poisonous gas. It forms two types of halides as PX_3 and PX_5. PCl_3 is prepared by the reaction of white phosphorus with dry chlorine while PCl_5 is prepared by the reaction of phosphorus with SO_2Cl_2. Phosphorus forms a number of oxoacids. Depending upon the number of P–OH groups, their basicity varies. The oxoacids which have P–H bonds are good reducing agents.

The Group 16 elements have general electronic configuration ns^2np^4. They show maximum oxidation state, +6. Gradation in physical and chemical properties is observed in the group 16 elements. In laboratory, dioxygen is prepared by heating $KClO_3$ in presence of MnO_2. It forms a number of oxides with metals. Allotropic form of oxygen is O_3 which is a highly oxidising agent. Sulphur forms a number of allotropes. Of these, α– and β– forms of sulphur are the most important. Sulphur combines with oxygen to give oxides such as SO_2 and SO_3. SO_2 is prepared by the direct union of sulphur with oxygen. SO_2 is used in the manufacture of H_2SO_4. Sulphur forms a number of oxoacids. Amongst them, most important is H_2SO_4. It is prepared by **contact process**. It is a dehydrating and oxidising agent. It is used in the manufacture of several compounds

Group 17 of the periodic table consists of the following elements F, Cl, Br, I and At.These elements are extremely reactive and as such they are found in the combined state only. The common oxidation state of these elements is –1. However, highest oxidation state can be +7. They show regular gradation in physical and chemical properties. They form oxides, hydrogen halides, interhalogen compounds and oxoacids. Chlorine is conveniently obtained by the reaction of HCl with $KMnO_4$. HCl is prepared by heating NaCl with concentrated H_2SO_4. Halogens combine with one another to form **interhalogen compounds** of the type $X\,X^1_n$ (n = 1, 3, 5, 7) where X^1 is lighter than X. A number of oxoacids of halogens are known. In the structures of these oxoacids, halogen is the central atom which is bonded in each case with one OH bond as X–OH. In some cases X = 0 bonds are also found

Group 18 of the periodic table consists of **noble gases**. They have $ns^2\ np^6$ valence shell electronic configuration except He which has $1s^2$. All the gases except Rn occur in atmosphere. Rn is obtained as the decay product of ^{226}Ra.

Due to complete octet of outermost shell, they have less tendency to form compounds. The best characterised compounds are those of xenon with fluorine and oxygen only under certain conditions. These gases have several uses. Argon is used to provide inert atmosphere, helium is used in filling balloons for meteorological observations, neon is used in discharge tubes and fluorescent bulbs.

IMPORTANT QUESTIONS (1M)

1) What is inert pair effect ?
2) Give one example each for normal oxide and mixed oxide of nitrogen?
3) What is allotrophy?
4) Why H_2O is a liquid while H_2S is a gas?
5) What is tailing of mercury ?
6) Write the uses of ozone?
7) What is aquaregia?
8) What are interhalogen compounds?
9) Which noble gas is radioactive?
10) Write the allotropic forms of phosphorus?

4 Marks

1) How ammonia is manufactured by Haber's process?
2) How is Nitric acid is manufactured by Ostwald's method?
3) Write the properties of ozone?
4) How is chlorine manufactured by Deacon's method?
5) Write the preparation and structures of xenon fluorides?

CHAPTER 7

THE d- AND f- BLOCK ELEMENTS

The *d*-block of the periodic table contains the elements of the groups 3-12 in which the *d* orbitals are progressively filled in each of the four long periods. The elements constituting the *f* -block are those in which the 4 *f* and 5 *f* orbitals are progressively filled in the latter two long periods; these elements are formal members of group 3 from which they have been taken out to form a separate *f*-block of the periodic table. The names *transition metals* and *inner transition* metals are often used to refer to the elements of *d*-and *f*-blocks respectively.

There are mainly three series of the transition metals, 3*d* series (Sc to Zn), 4*d* series (Y to Cd) and 5*d* series (La to Hg, omitting Ce to Lu). The fourth 6*d* series which begins with Ac is still incomplete. The two series of the inner transition metals, (4*f* and 5*f*) are known as *lanthanoids* and *actinoids* respectively.

Strictly speaking, a transition element is defined as the one which has incompletely filled *d* orbitals in its ground state or in any one of its oxidation states. Zinc, cadmium and mercury of group 12 have full *d*10 configuration in their ground state as well as in their common oxidation states and hence, are not regarded as transition metals. However, being the end members of the three transition series, their chemistry is studied along with the chemistry of the transition metals.

The presence of partly filled *d* or *f* orbitals in their atoms sets the study of the transition elements and their compounds apart from that of the main group elements. However, the usual theory of valence as applicable to the main group elements can also be applied successfully to the transition elements.

Various precious metals such as silver, gold and platinum and industrially important metals like iron, copper and titanium form part of the transition metals.

In this Unit, besides introduction, we shall first deal with the electronic configuration, occurrence and general characteristics of the transition elements with special emphasis on the trends in the properties of the first row (3*d*) transition metals and the preparation and properties of some important compounds. This will be followed by consideration of certain general aspects such as electronic configurations, oxidation states and chemical reactivity of the inner transition metals.

THE TRANSITION ELEMENTS (*d*-BLOCK)

7.1 Position in the Periodic Table

The *d*–block occupies the large middle section flanked by *s*– and *p*– blocks in the periodic table. The very name 'transition' given to the elements of *d*-block is only because of their position between *s*– and *p*– block elements. The *d*–orbitals of the penultimate energy level in their atoms receive electrons giving rise to the three rows of the transition metals, i.e., 3*d*, 4*d* and 5*d*. The fourth row of 6*d* is still incomplete. These series of the transition elements are shown in the following table.

7.2 Electronic Configurations of the d-Block Elements

In general the electronic configuration of these elements is (n-1) d^{1-10}ns$^{1-2}$. The (n–1) stands for the inner *d* orbitals which may have one to ten electrons and the outermost n*s* orbital may have one or two electrons. However, this generalization has several exceptions because of very little energy difference between (n-1)*d* and n*s* orbitals. Furthermore, half and completely filled sets of orbitals are relatively more stable. A consequence of this factor is reflected in the

electronic configurations of Cr and Cu in the 3*d* series. Consider the case of Cr, for example, which has $3d^5\ 4s^1$ instead of $3d^4 4s^2$; the energy gap between the two sets (3*d* and 4*s*) of orbitals is small enough to prevent electron entering the 3*d* orbitals. Similarly in case of Cu, the configuration is $3d^{10}4s^1$ and not $3d^9 4s^2$. The outer electronic configurations of the transition elements are given in Table. 7.1

Table 7.1 Outer Electronic Configurations of the Transition Elements (ground state)

1st Series										
	Sc	T i	V	Cr	Mn	Fe	Co	Ni	Cu	Zn
Z	21	22	23	24	25	26	27	28	29	30
4*s*	2	2	2	1	2	2	2	2	1	2
3*d*	1	2	3	5	5	6	7	8	10	10

2nd Series										
	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd
Z	39	40	41	42	43	44	45	46	47	48
5*s*	2	2	1	1	1	1	1	0	1	2
4*d*	1	2	4	5	6	7	8	10	10	10

3rd Series										
	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg
Z	57	72	73	74	75	76	77	78	79	80
6*s*	2	2	2	2	2	2	2	1	1	2
5*d*	1	2	3	4	5	6	7	9	10	10

4th Series										
	Ac	Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg	Uub
Z	89	104	105	106	107	108	109	110	111	112
7*s*	2	2	2	2	2	2	2	2	1	2
6*d*	1	2	3	4	5	6	7	8	10	10

The electronic configurations of Zn, Cd and Hg are represented by the general formula $(n\text{-}1)d^{10}ns^2$. The orbitals in these elements are completely filled in the ground state as well as in their common oxidation states. Therefore, they are not regarded as transition elements.

The *d* orbitals of the transition elements project to the periphery of an atom more than the other orbitals (i.e., *s* and *p*), hence, they are more influenced by the surroundings as well as affecting the atoms or molecules surrounding them. In some respects, ions of a given d^n configuration (n=1–9) have similar magnetic and electronic properties. With partly filled *d* orbitals these elements exhibit certain characteristic properties such as display of a variety of oxidation states, formation of coloured ions and entering into complex formation with a variety of ligands.

The transition metals and their compounds also exhibit catalytic property and paramagnetic behaviour. All these characteristics have been discussed in detail later in this Unit.

There are greater horizontal similarities in the properties of the transition elements in contrast to the main group elements. However, some group similarities also exist. We shall first study the general characteristics and their trends in the horizontal rows (particularly 3*d* row) and then consider some group similarities.

7.3 General Properties of the Transition Elements (d-Block)

7.3.1 Physical Properties

Nearly all the transition elements display typical metallic properties such as high tensile strength, ductility malleability, high thermal and electrical conductivity and metallic luster. With the exceptions of Zn, Cd, Hg and Mn, they have one or more typical metallic structures at normal temperatures.

Lattice Structures of Transition Metals

Sc	**Ti**	**V**	**Cr**	**Mn**	**Fe**	**Co**	**Ni**	**Cu**	**Zn**
Hcp (bcc)	hcp (bcc)	bcc	bcc (bcc, ccp)	X (hcp)	bcc (hcp)	ccp	ccp	ccp	X (hcp)
Y	**Zr**	**Nb**	**Mo**	**Tc**	**Ru**	**Rh**	**Pd**	**Ag**	**Cd**
Hcp (bcc)	hcp (bcc)	bcc	bcc	hcp	hcp	ccp	ccp	ccp	X (hcp)
La	**Hf**	**Ta**	**W**	**Re**	**Os**	**Ir**	**Pt**	**Au**	**Hg**
Hcp (ccp,bcc)	hcp (bcc)	bcc	bcc	hcp	hcp	ccp	ccp	ccp	X

(bcc = body centred cubic; hcp = hexagonal close packed; ccp = cubic close packed; X = a typical metal structure).

The transition metals (with the exception of Zn, Cd and Hg) are very much hard and have low volatility. Their melting and boiling points are high. Fig. 7.1 depicts the melting points of the 3*d*, 4*d* and 5*d* transition metals. The high melting points of these metals are attributed to the involvement of greater number of electrons from (n-1)*d* in addition to the ns electrons in the inter atomic metallic bonding. In any row the melting points of these metals rise to a maximum at d^5 except for anomalous values of Mn and Tc and fall regularly as the atomic number increases. They have high enthalpies of atomization which are shown in the below Fig. 7.2 The maxima at about the middle of each series indicate that one unpaired electron per *d* orbital is particularly favourable for strong interatomic interaction. In general, greater the number of valence electrons, stronger is the resultant bonding. Since the enthalpy of atomization is an important factor in determining the standard electrode potential of a metal, metals with very high enthalpy of atomization (i.e., very high boiling point) tend to be noble in their reactions (see later for electrode potentials).

Another generalization that may be drawn from Fig. 7.2 is that the metals of the second and third series have greater enthalpies of atomization than the corresponding elements of the

first series; this is an important factor in accounting for the occurrence of much more frequent metal – metal bonding in compounds of the heavy transition metals.

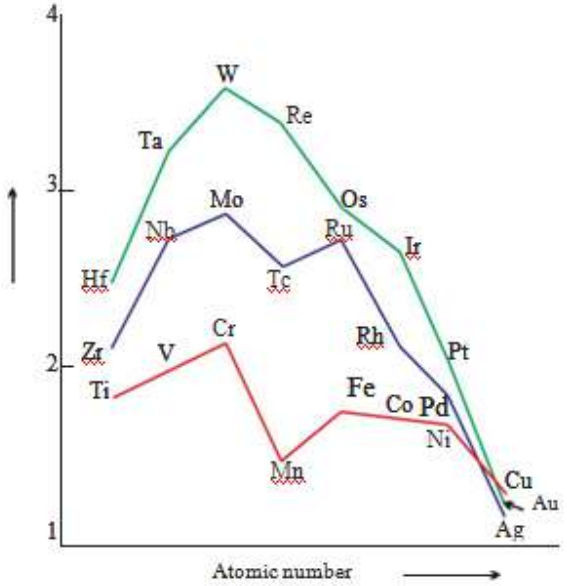


Fig. 7.1 Trends in melting points of transition elements

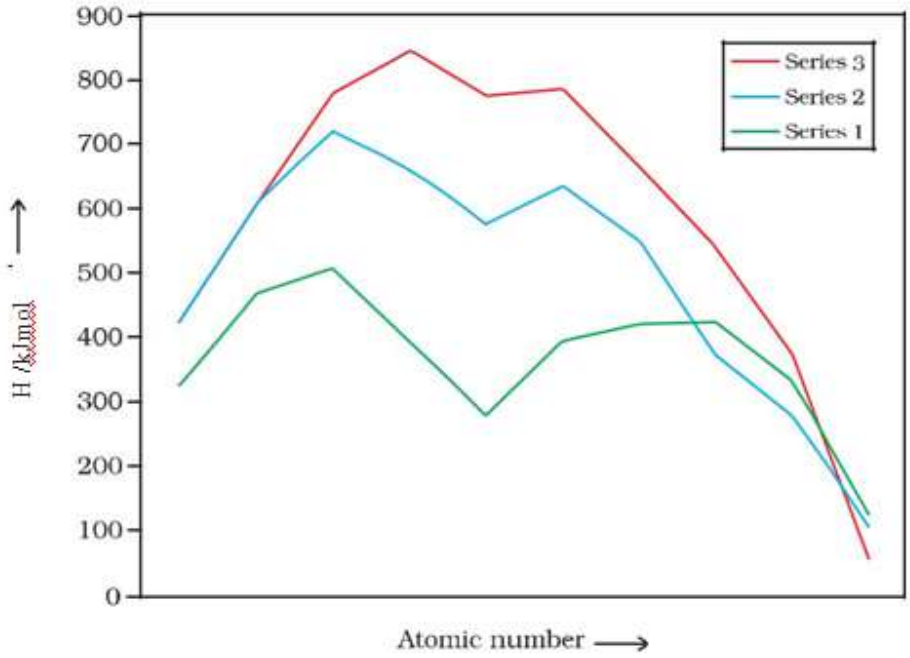


Fig. 7.2 Trends in enthalpies of atomization of transition elements

7.3.2 Variation in Atomic and Ionic Sizes of Transition Metals

In general, ions of the same charge in a given series show progressive decrease in radius with increasing atomic number. This is because the new electron enters a *d* orbital each time the nuclear charge increases by unity. It may be recalled that the shielding effect of a *d* electron is not that effective, hence the net electrostatic attraction between the nuclear charge and the outermost electron increases and the ionic radius decreases. The same trend is observed in the atomic radii of a given series. However, the variation within a series is quite small. An interesting point emerges when atomic sizes of one series are compared with those of the corresponding elements in the other series. The curves in Fig. 7.3 show an increase from the first (3*d*) to the second (4*d*) series of the elements but the radii of the third (5*d*) series are virtually the same as those of the corresponding members of the second series. This phenomenon is associated with the intervention of the 4*f* orbitals which must be filled before the 5*d* series of elements begin. The filling of 4*f* before 5*d* orbital results in a regular decrease in atomic radii called **Lanthanoid contraction** which essentially compensates for the expected increase in atomic size with increasing atomic number. The net result of the lanthanoid contraction is that the second and the third *d* series exhibit similar radii (e.g., Zr 160 pm, Hf 159 pm) and have very similar physical and chemical properties much more than that expected on the basis of usual family relationship.

The factor responsible for the lanthanoid contraction is somewhat similar to that observed in an ordinary transition series and is attributed to similar cause, i.e., the imperfect shielding of one electron by another in the same set of orbitals. However, the shielding of one 4*f* electron by another is less than that of one *d* electron by another, and as the nuclear charge increases along the series, there is fairly regular decrease in the size of the entire 4*f* n orbitals.

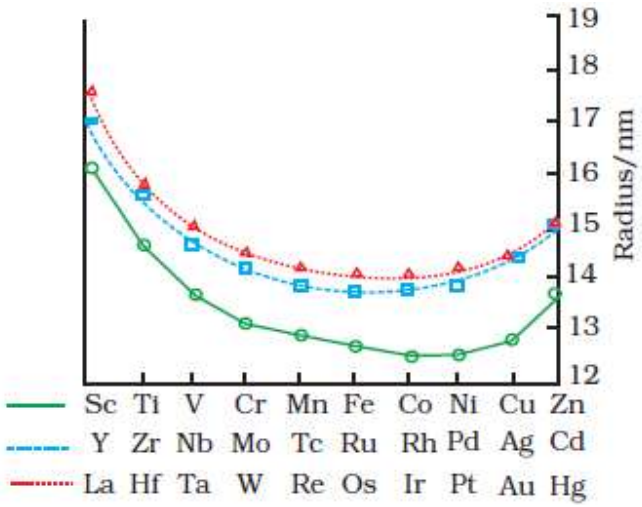


Fig. 7.3 Trends in atomic radii of transition elements

Table 7.2 Electronic Configurations and some other Properties of the First Series of Transition Elements

Element		Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn
Atomic number		21	22	23	24	25	26	27	28	29	30
Electronic configuration											
	M	$3d^14s^2$	$3d^24s^2$	$3d^34s^2$	$3d^54s^1$	$3d^54s^2$	$3d^64s^2$	$3d^74s^2$	$3d^84s^2$	$3d^{10}4s^1$	$3d^{10}4s^2$
	M^+	$3d^14s^1$	$3d^24s^1$	$3d^34s^1$	$3d^5$	$3d^54s^1$	$3d^64s^1$	$3d^74s^1$	$3d^84s^1$	$3d^{10}$	$3d^{10}4s^1$
	M^{2+}	$3d^1$	$3d^2$	$3d^3$	$3d^4$	$3d^5$	$3d^6$	$3d^7$	$3d^8$	$3d^9$	$3d^{10}$
	M^{3+}	[Ar]	$3d^1$	$3d^2$	$3d^3$	$3d^4$	$3d^5$	$3d^6$	$3d^7$	-	-
Enthalpy of atomisation, $\Delta_a H^\ominus$/kJ mol^{-1}		326	473	515	397	281	416	425	430	339	126
Ionisation enthalpy/ $\Delta_i H^\ominus$/kJ mol^{-1}											
$\Delta_i H_1^\ominus$	I	631	656	650	653	717	762	758	736	745	906
$\Delta_i H_2^\ominus$	II	1235	1309	1414	1592	1509	1561	1644	1752	1958	1734
$\Delta_i H_3^\ominus$	III	2393	2657	2833	2990	3260	2962	3243	3402	3556	3829
Metallic/ionic radii/pm	M	164	147	135	129	137	126	125	125	128	137
	M^{2+}	-	-	79	82	82	77	74	70	73	75
	M^{3+}	73	67	64	62	65	65	61	60	-	-
Standard electrode potential $E^\ominus$/V	M^{2+}/M	-	–1.63	–1.18	–0.90	–1.18	–0.44	–0.28	–0.25	+0.34	-0.76
	M^{3+}/M^{2+}	-	–0.37	–0.26	–0.41	+1.57	+0.77	+1.97	-	-	-
Density/g cm^{-3}		3.43	4.1	6.07	7.19	7.21	7.8	8.7	8.9	8.9	7.1

7.3.3 Ionisation Enthalpies

Due to an increase in nuclear charge which accompanies the filling of the inner *d* orbitals, there is an increase in ionisation enthalpy along each series of the transition elements from left to right. However, many small variations occur. Table 7.2 gives the values for the first three ionisation enthalpies of the first row elements. These values show that the successive enthalpies of these elements do not increase as steeply as in the main group elements. Although the first ionisation enthalpy, in general, increases, the magnitude of the increase in the second and third ionisation enthalpies for the successive elements, in general, is much higher.

The irregular trend in the first ionization enthalpy of the 3*d* metals, though of little chemical significance, can be accounted for by considering that the removal of one electron alters the relative energies of 4*s* and 3*d* orbitals. So the unipositive ions have d^n configurations with no 4*s* electrons. There is thus, a reorganization energy accompanying ionization with some gains in exchange energy as the number of electrons increases and from the transference of *s* electrons into *d* orbitals. There is the generally expected increasing trend in the values as the effective nuclear charge increases. However, the value of Cr is lower because of the absence of any change in the *d* configuration and the value for Zn higher because it represents an ionization from the 4*s* level. The lowest common oxidation state of these metals is +2. To form the M^{2+} ions from the gaseous atoms, the sum of the first and second ionization energies is required in addition to the enthalpy of atomization for each element. The dominant term is the second ionization enthalpy which shows unusually high values for Cr and Cu where the d^5 and d^{10} configurations of the M^+ ions are disrupted, with considerable loss of exchange energy. The

value for Zn is correspondingly low as the ionization consists of the removal of an electron which allows the production of the stable d^{10} configuration. The trend in the third ionization enthalpies is not complicated by the 4*s* orbital factor and shows the greater difficulty of removing an electron from the d^5 (Mn^{2+}) and d^{10} (Zn^{2+}) ions superimposed upon the general increasing trend. In general, the third ionization enthalpies are quite high and there is a marked break between the values for Mn^{2+} and Fe^{2+}. Also the high values for copper, nickel and zinc indicate why it is difficult to obtain oxidation state greater than two for these elements.

` Although ionization enthalpies give some guidance concerning the relative stabilities of oxidation states, this problem is very complex and not amendable to ready generalisation.

7.3.4 Oxidation States

One of the notable features of a transition element is the great variety of oxidation states it may show in its compounds. Table 7.3 lists the common oxidation states of the first row transition elements.

Table 7.3 Oxidation States of the first row Transition Metals (the most common ones are in bold types)

Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn
	+2	+2	+2	**+2**	**+2**	**+2**	**+2**	+1	**+2**
+3	+3	+3	**+3**	+3	**+3**	**+3**	+3	**+2**	
	+4	+4	+4	+4	+4	+4	+4		
		+5	+5	+5					
			+6	+6	+6				
				+7					

The elements which give the greatest number of oxidation states occur in or near the middle of the series. Manganese, for example, exhibits all the oxidation states from +2 to +7. The lesser number of oxidation states at the extreme ends stems from either too few electrons to lose or share (Sc, Ti) or too many d electrons (hence fewer orbitals available in which to share electrons with others) for higher valence (Cu, Zn). Thus, early in the series scandium(II) is virtually unknown and titanium (IV) is more stable than Ti(III) or Ti(II). At the other end, the only oxidation state of zinc is +2 (no *d* electrons are involved). The maximum oxidation states of reasonable stability correspond in value to the sum of the *s* and *d* electrons upto manganese ($Ti^{IV}O_2$, $V^{V}O_2^{+}$, $Cr^{V1}O_4^{2-}$, $Mn^{VII}O^{4-}$) followed by a rather abrupt decrease in stability of higher oxidation states, so that the typical species to follow are $Fe^{II, III}$, $Co^{II, III}$, Ni^{II}, $Cu^{I, II}$, Zn^{II}.

The variability of oxidation states, a characteristic of transition elements, arises out of incomplete filling of *d* orbitals in such a way that their oxidation states differ from each other by unity, e.g., V^{II}, V^{III}, V^{IV}, V^{V}. This is in contrast with the variability of oxidation states of non transition elements where oxidation states normally differ by a unit of two.

An interesting feature in the variability of oxidation states of the *d*–block elements is noticed among the groups (groups 4 through 10). Although in the *p*–block the lower oxidation states are favoured by the heavier members (due to inert pair effect), the opposite is true in the groups of *d*-block. For example, in group 6, Mo(VI) and W(VI) are found to be more stable than Cr(VI). Thus Cr(VI) in the form of dichromate in acidic medium is a strong oxidising agent, whereas MoO_3 and WO_3 are not.

Low oxidation states are found when a complex compound has ligands capable of π-acceptor character in addition to the σ-bonding. For example, in $Ni(CO)_4$ and $Fe(CO)_5$, the oxidation state of nickel and iron is zero.

7.3.5 Trends in Stability of Higher Oxidation States

The highest oxidation numbers are achieved in TiX_4 (tetrahalides), VF_5 and CrF_6. The +7 state for Mn is not represented in simple halides but MnO_3F is known, and beyond Mn no metal has a trihalide except FeX_3 and CoF_3. The ability of fluorine to stabilise the highest oxidation state is due to either higher lattice energy as in the case of CoF_3, or higher bond enthalpy terms for the higher covalent compounds, e.g., VF_5 and CrF_6. Although VV is represented only by VF_5, the other halides, however, undergo hydrolysis to give oxohalides, VOX_3. Another feature of fluorides is their instability in the low oxidation states e.g., VX_2 (X = CI, Br or I) and the same applies to CuX. On the other hand, all Cu^{II} halides are known except the iodide. In this case, Cu^{2+} oxidises I^- to I_2.

$$2Cu^{2+} + 4I^- \rightarrow Cu_2I_2(s) + I_2$$

However, many copper (I) compounds are unstable in aqueous solution and undergo disproportionation.

$$2Cu^+ \rightarrow Cu^{2+} + Cu$$

The stability of Cu^{2+} (aq) rather than Cu^+(aq) is due to the much more negative $\Delta_{hyd}H^{\ominus}$ of Cu^{2+} (aq) than Cu+, which more than compensates for the second ionisation enthalpy of Cu. The ability of oxygen to stabilise the highest oxidation state is demonstrated in the oxides. The highest oxidation number in the oxides coincides with the group number and is attained in Sc_2O_3 to Mn_2O_7. Beyond Group 7, no higher oxides of Fe above Fe_2O_3, are known, although ferrates (VI) $(FeO_4)^{2-}$, are formed in alkaline media but they readily decompose to Fe_2O_3 and O_2. Besides the oxides, oxocations stabilise V^v as VO^{2+}, V^{IV} as VO^{2+} and Ti^{IV} as TiO^{2+}. The ability of oxygen to stabilise these high oxidation states exceeds that of fluorine. Thus the highest Mn fluoride is MnF_4 whereas the highest oxide is Mn_2O_7. The ability of oxygen to form multiple bonds to metals explains its superiority. In the covalent oxide Mn_2O_7, each Mn is tetrahedrally surrounded by O's including a Mn–O–Mn bridge. The tetrahedral $[MO_4]^{n-}$ ions are known for V^V, Cr^{VI}, Mn^V, Mn^{VI} and Mn^{VII}.

Table 7.4 Oxides of 3d metal ions

Oxidation Number	Groups									
	3	4	5	6	7	8	9	10	11	12
+ 7					Mn_2O_7					
+ 6				CrO_3						
+ 5			V_2O_5							
+ 4		TiO_2	V_2O_4	CrO_2	MnO_2					
+ 3	Sc_2O_3	Ti_2O_3	V_2O_3	Cr_2O_3	Mn_2O_3	Fe_2O_3				
					$Mn_3O_4^*$	$Fe_3O_4^*$	$Co_3O_4^*$			
+ 2		TiO	VO	(CrO)	MnO	FeO	CoO	NiO	CuO	ZnO
+ 1									Cu_2O	

7.3.6 Magnetic Properties

When a magnetic field is applied to substances, mainly two types of magnetic behaviour are observed: *diamagnetism* and *paramagnetism* . Diamagnetic substances are repelled by the applied field while the paramagnetic substances are attracted. Substances which are attracted very strongly are said to be *ferromagnetic*. In fact, ferromagnetism is an extreme form of paramagnetism. Many of the transition metal ions are paramagnetic. Paramagnetism arises from the presence of unpaired electrons, each such electron having a magnetic moment associated with its spin angular momentum and orbital angular momentum. For the compounds of the first series of transition metals, the contribution of the orbital angular momentum is effectively quenched and hence is of no significance. For these, the magnetic moment is determined by the number of unpaired electrons and is calculated by using the 'spin-only' formula, i.e., $\mu = n(n + 2)$ where n is the number of unpaired electrons and μ is the magnetic moment in units of Bohr magneton (BM). A single unpaired electron has a magnetic moment of 1.73 Bohr magnetons (BM). The magnetic moment increases with the increasing number of unpaired electrons. Thus, the observed magnetic moment gives a useful indication about the number of unpaired electrons present in the atom, molecule or ion. The magnetic moments calculated from the 'spin-only' formula and those derived experimentally for some ions of the first row transition elements are given in Table 7.5. The experimental data are mainly for hydrated ions in solution or in the solid state.

Table 7.5 Calculated and Observed Magnetic Moments (BM)

Ion	Configuration	Unpaired electron(s)	Magnetic moment	
			Calculated	Observed
Sc^{3+}	$3d^0$	0	0	0
Ti^{3+}	$3d^1$	1	1.73	1.75
Tl^{2+}	$3d^2$	2	2.84	2.76
V^{2+}	$3d^3$	3	3.87	3.86
Cr^{2+}	$3d^4$	4	4.90	4.80
Mn^{2+}	$3d^5$	5	5.92	5.96
Fe^{2+}	$3d^6$	4	4.90	5.3 – 5.5
Co^{2+}	$3d^7$	3	3.87	4.4 – 5.2
Ni^{2+}	$3d^8$	2	2.84	2.9 – 3, 4
Cu^{2+}	$3d^9$	1	1.73	1.8 – 2.2
Zn^{2+}	$3d^{10}$	0	0	

7.3.7 Formation of coloured ions

When an electron from a lower energy *d* orbital is excited to a higher energy *d* orbital, the energy of excitation corresponds to the frequency of light absorbed (Unit 9). This frequency generally lies in the visible region. The colour observed corresponds to the complementary colourof the light absorbed. The frequency of the light absorbed is determined by the nature of the ligand. In aqueous solutions where water molecules are the ligands, the colours of the ions

observed are listed in Table 8.8. A few coloured solutions of *d*–block elements are illustrated in Fig 7.4

Fig 7.4 Colours of the first row transition metal ions in aqueous solutions. From left to right V^{4+}, Mn^{2+}, Fe^{3+},Ni^{2+} and Cu^{2+}

7.3.8 Formation of Complex Compounds

Complex compounds are those in which the metal ions bind a number of anions or neutral molecules giving complex species with characteristic properties. A few examples are: $[Fe(CN)_6]^{3-}$, $[Fe(CN)_6]^{4-}$, $[Cu(NH^3)^4]^{2+}$ and $[PtCl_4]^{2-}$. (The chemistry of complex compounds is dealt with in detail in Unit 9). The transition metals form a large number of complex compounds. This is due to the comparatively smaller sizes of the metal ions, their high ionic charges and the availability of *d* orbitals for bond formation.

7.3.9 Catalytic Properties

The transition metals and their compounds are known for their catalytic activity. This activity is ascribed to their ability to adopt multiple oxidation states and to form complexes. Vanadium (V) oxide (in Contact Process), finely divided iron (in Haber's Process), and nickel (in Catalytic Hydrogenation) are some of the examples. Catalysts at a solid surface involve the formation of bonds between reactant molecules and atoms of the surface of the catalyst (first row transition metals utilize 3*d* and 4*s* electrons for bonding). This has the effect of increasing the concentration of the reactants at the catalyst surface and also weakening of the bonds in the reacting molecules (the activation energy is lowering). Also because the transition metal ions can change their oxidation states, they become more effective as catalysts. For example, iron (III) catalyses the reaction between iodide and per sulphate ions.

$$2\ I^- + S_2O_8^{2-} \rightarrow I_2 + 2\ SO_4^{2-}$$

An explanation of this catalytic action can be given as:

$$2\ Fe^{3+} + 2\ I^- \rightarrow 2\ Fe^{2+} + I_2$$

$$2\ Fe^{2+} + S_2O_8^{2-} \rightarrow 2\ Fe^{3+} + 2SO_4^{2-}$$

7.3.10 Formation of Interstitial Compounds

Interstitial compounds are those which are formed when small atoms like H, C or N are trapped inside the crystal lattices of metals. They are usually non stoichiometric and are neither typically ionic nor covalent for example, TiC, Mn_4N, Fe_3H, $VH_{0.56}$ and $TiH_{1.7}$, etc. The formulas quoted do not, of course, correspond to any normal oxidation state of the metal. Because of the nature of their composition, these compounds are referred to as *interstitial* compounds. The principal physical and chemical characteristics of these compounds are as follows:

(i) They have high melting points, higher than those of pure metals.

(ii) They are very hard, some borides approach diamond in hardness.

(iii) They retain metallic conductivity.

(iv) They are chemically inert

7.3.11 Alloy Formation

An alloy is a blend of metals prepared by mixing the components. Alloys may be homogeneous solid solutions in which the atoms of one metal are distributed randomly among the atoms of the other. Such alloys are formed by atoms with metallic radii that are within about 15 percent of each other. Because of similar radii and other characteristics of transition metals, alloys are readily formed by these metals. The alloys so formed are hard and have often high melting points. The best known are ferrous alloys: chromium, vanadium, tungsten, molybdenum and manganese are used for the production of a variety of steels and stainless steel. Alloys of transition metals with non transition metals such as brass (copper-zinc) and bronze (copper-tin), are also of considerable industrial importance

7.4 THE INNER TRANSITION ELEMENTS (*f*-BLOCK)

The *f*-block consists of the two series, lanthanoids (the fourteen elements following lanthanum) and actinoids (the fourteen elements following actinium). Because lanthanum closely resembles the lanthanoids, it is usually included in any discussion of the lanthanoids for which the general symbol Ln is often used. Similarly, a discussion of the actinoids includes actinium besides the fourteen elements constituting the series.

The lanthanoids resemble one another more closely than do the members of ordinary transition elements in any series. They have only one stable oxidation state and their chemistry provides an excellent opportunity to examine the effect of small changes in size and nuclear charge along a series of otherwise similar elements.

The chemistry of the actinoids is, on the other hand, much more complicated. The complication arises partly owing to the occurrence of a wide range of oxidation states in these elements and partly because their radioactivity creates special problems in their study; the two series will be considered separately here.

7.4.1 The Lanthanoids

The names, symbols, electronic configurations of atomic and some ionic states and atomic and ionic radii of lanthanum and lanthanoids (for which the general symbol Ln is used)

7.4.2 Electronic Configurations

It may be noted that atoms of these elements have electronic configuration with 6*s*2 common but with variable occupancy of 4*f* level. However, the electronic configurations of all the tri positive ions (the most stable oxidation state of all the lanthanoids) are of the form 4*f* n (n = 1 to 14 with increasing atomic number).

7.4.3 Atomic and Ionic Sizes

The overall decrease in atomic and ionic radii from lanthanum to lutetium (the **lanthanoid contraction**) is a unique feature in the chemistry of the lanthanoids. It has far reaching consequences in the chemistry of the third transition series of the elements. The decrease in atomic radii (derived from the structures of metals) is not quite regular as it is regular in M^{3+} ions. This contraction is, of course, similar to that observed in an ordinary transition series and is attributed to the same cause, the imperfect shielding of one electron by another in the same sub-shell. However, the shielding of one 4 *f* electron by another is less than one *d* electron by another with the increase in nuclear charge along the series. There is fairly regular

decrease in the sizes with increasing atomic number. The cumulative effect of the contraction of the lanthanoid series, known as *lanthanoid contraction*, causes the radii of the members of the third transition series to be very similar to those of the corresponding members of the second series. The almost identical radii of Zr (160 pm) and Hf (159 pm), a consequence of the lanthanoid contraction, account for their occurrence together in nature and for the difficulty faced in their separation.

7.4.4 Oxidation States

. In the lanthanoids, La(III) and Ln(III) compounds are predominant species. However, occasionally +2 and +4 ions in solution or in solid compounds are also obtained. This irregularity (as in ionisation enthalpies) arises mainly from the extra stability of empty, half-filled or filled *f* subshell. Thus, the formation of Ce^{IV} is favoured by its noble gas configuration, but it is a strong oxidant reverting to the common +3 state. The E^o value for Ce^{4+}/ Ce^{3+} is +1.74 V which suggests that it can oxidise water. However, the reaction rate is very slow and hence Ce(IV) is a good analytical reagent. Pr, Nd, Tb and Dy also exhibit +4 state but only in oxides, MO_2. Eu^{2+} is formed by losing the two *s* electrons and its f^7 configuration accounts for the formation of this ion. However, Eu^{2+} is a strong reducing agent changing to the common +3 state. Similarly Yb^{2+} which has f^{14} configuration is a reductant. Tb^{IV} has half-filled *f*-orbitals and is an oxidant. The behaviour of samarium is very much like europium, exhibiting both +2 and +3 oxidation states.

Table 7.6 Electronic Configurations and Radii of Lanthanum and Lanthanoids

Atomic Number	Name	Symbol	Electronic configurations*			Radii/pm		
			Ln	Ln^{2+}	Ln^{3+}	Ln^{4+}	Ln	Ln^{3+}
57	Lanthanum	La	$5d^16s^2$	$5d^1$	$4f^0$		187	106
58	Cerium	Ce	$4f^15d^16s^2$	$4f^2$	$4f^1$	$4f^0$	183	103
59	Praseodymium	Pr	$4f^36s^2$	$4f^3$	$4f^2$	$4f^1$	182	101
60	Neodymium	Nd	$4f^46s^2$	$4f^4$	$4f^3$	$4f^2$	181	99
61	Promethium	Pm	$4f^56s^2$	$4f^5$	$4f^4$		181	98
62	Samarium	Sm	$4f^66s^2$	$4f^6$	$4f^5$		180	96
63	Europium	Eu	$4f^76s^2$	$4f^7$	$4f^6$		199	95
64	Gadolinium	Gd	$4f^75d^16s^2$	$4f^75d^1$	$4f^7$		180	94
65	Terbium	Tb	$4f^96s^2$	$4f^9$	$4f^8$	$4f^7$	178	92
66	Dysprosium	Dy	$4f^{10}6s^2$	$4f^{10}$	$4f^9$	$4f^8$	177	91
67	Holmium	Ho	$4f^{11}6s^2$	$4f^{11}$	$4f^{10}$		176	89
68	Erbium	Er	$4f^{12}6s^2$	$4f^{12}$	$4f^{11}$		175	88
69	Thulium	Tm	$4f^{13}6s^2$	$4f^{13}$	$4f^{12}$		174	87
70	Ytterbium	Yb	$4f^{14}6s^2$	$4f^{14}$	$4f^{13}$		173	86
71	Lutetium	Lu	$4f^{14}5d^16s^2$	$4f^{14}5d^1$	$4f^{14}$	–	–	–

7.4.5 General Characteristics

All the lanthanoids are silvery white soft metals and tarnish rapidly in air. The hardness increases with increasing atomic number, samarium being steel hard. Their melting points range between 1000 to 1200 K but samarium melts at 1623 K. They have typical metallic structure and are good conductors of heat and electricity. Density and other properties change smoothly except for Eu and Yb and occasionally for Sm and Tm. Many trivalent lanthanoid ions are coloured both in the solid state and in aqueous solutions. Colour of these ions may be attributed to the presence of *f* electrons. Neither La^{3+} nor Lu^{3+} ion shows any colour but the rest do so. However, absorption bands are narrow, probably because of the excitation within *f* level. The lanthanoid ions other than the f^0 type (La^{3+} and Ce^{4+}) and the f^{14} type (Yb^{2+} and Lu^{3+}) are all paramagnetic. The paramagnetism rises to maximum in neodymium. The first ionisation enthalpies of the lanthanoids are around 600 kJ mol^{-1}, the second about 1200 kJ mol^{-1} comparable with those of calcium. A detailed discussion of the variation of the third ionisation enthalpies indicates that the exchange enthalpy considerations (as in 3*d* orbitals of the first transition series), appear to impart a certain degree of stability to empty, half-filled and completely filled orbitals *f* level. This is indicated from the abnormally low value of the third ionisation enthalpy of lanthanum, gadolinium and lutetium. In their chemical behaviour, in general, the earlier members of the series are quite reactive similar to calcium but, with increasing atomic number, they behave more like aluminium. Values for *E*V for the half-reaction

$$Ln^{3+}(aq) + 3e^- \rightarrow Ln(s)$$

are in the range of –2.2 to –2.4 V except for Eu for which the value is – 2.0 V. This is, of course, a small variation. The metals combine with hydrogen when gently heated in the gas. The carbides, Ln3C, Ln2C3 and LnC2 are formed when the metals are heated with carbon. They liberate hydrogen from dilute acids and burn in halogens to form halides. They form oxides M2O3 and hydroxides M(OH)3. The hydroxides are definite compounds, not just hydrated oxides. They are basic like alkaline earth metal oxides and hydroxides.

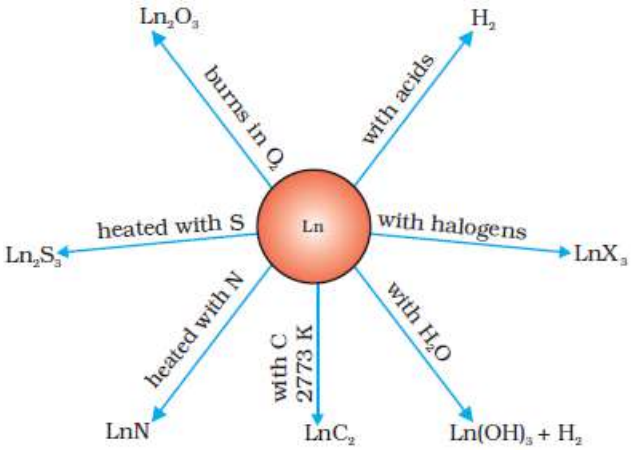


Fig 7.5 Chemical reactions of the lanthanoids

The best single use of the lanthanoids is for the production of alloy steels for plates and pipes. A well known alloy is *mischmetall* which consists of a lanthanoid metal (~ 95%) and iron (~ 5%) and traces of S, C, Ca and Al. A good deal of mischmetall is used in Mg-based alloy to produce bullets, shell and lighter flint. Mixed oxides of lanthanoids are employed as catalysts in

petroleum cracking. Some individual Ln oxides are used as phosphorsin television screens and similar fluorescing surfaces.

7.5 The Actinoids

The actinoids include the fourteen elements from Th to Lr. The names, symbols and some properties of these elements are given the following table.

Table 7.7 Some properties of Actinium and Actinoids

Atomic Number	Name	Symbol	Electronic conifigurations*		Radii/pm		
			M	M^{3+}	M^{4+}	M^{3+}	M^{4+}
89	Actinium	Ac	$6d^17s^2$	$5f^0$		111	
90	Thorium	Th	$6d^27s^2$	$5f^1$	$5f^0$		99
91	Protactinium	Pa	$5f^26d^17s^2$	$5f^2$	$5f^1$		96
92	Uranium	U	$5f^36d^17s^2$	$5f^3$	$5f^2$	103	93
93	Neptunium	Np	$5f^46d^17s^2$	$5f^4$	$5f^3$	101	92
94	Plutonium	Pu	$5f^67s^2$	$5f^5$	$5f^4$	100	90
95	Americium	Am	$5f^77s^2$	$5f^6$	$5f^5$	99	89
96	Curium	Cm	$5f^76d^17s^2$	$5f^7$	$5f^7$	99	88
97	Berkelium	Bk	$5f^97s^2$	$5f^8$	$5f^7$	98	87
98	Californium	Cf	$5f^{10}7s^2$	$5f^9$	$5f^8$	98	86
99	Einstenium	Es	$5f^{11}7s^2$	$5f^{10}$	$5f^9$	–	–
100	Fermium	Fm	$5f^{12}7s^2$	$5f^{11}$	$5f^{10}$	–	–
101	Mendelevium	Md	$5f^{13}7s^2$	$5f^{12}$	$5f^{11}$	–	–
102	Nobelium	No	$5f^{14}7s^2$	$5f^{13}$	$5f^{12}$	–	–
103	Lawrencium	Lr	$5f^{14}6d^17s^2$	$5f^{14}$	$5f^{13}$	–	–

The actinoids are radioactive elements and the earlier members have relatively long half-lives, the latter ones have half-life values ranging from a day to 3 minutes for lawrencium (Z =103). The latter members could be prepared only in nano gram quantities. These facts render their study more difficult.

7.5.1 Electronic Configurations

All the actinoids are believed to have the electronic configuration of 7s2 and variable occupancy of the 5f and 6d subshells. The fourteen electrons are formally added to 5*f*, though not in thorium (Z = 90) but from Pa onwards the 5*f* orbitals are complete at element 103. The irregularities in the electronic configurations of the actinoids, like those in the lanthanoids are related to the stabilities of the f^0, f^7 and f^{14} occupancies of the 5f orbitals. Thus, the configurations of Am and Cm are [Rn] 5*f* 77*s*2 and [Rn] $5f^76d^17s^2$. Although the 5*f* orbitals resemble the 4*f* orbitals in their angular part of the wave-function, they are not as buried as 4*f* orbitals and hence 5*f* electrons can participate in bonding to a far greater extent.

7.5.2 Ionic Sizes

The general trend in lanthanoids is observable in the actinoids as well. There is a gradual decrease in the size of atoms or M^{3+} ions across the series. This may be referred to as the *actinoid contraction* (like lanthanoid contraction). The contraction is, however, greater from element to element in this series resulting from poor shielding by 5*f* electrons.

Table 7.8 Electronic Configurations and Radii of Lanthanum and Lanthanoids

Atomic Number	Name	Symbol	Electronic configurations*			Radii/pm		
			Ln	Ln^{2+}	Ln^{3+}	Ln^{4+}	Ln	Ln^{3+}
57	Lanthanum	La	$5d^1 6s^2$	$5d^1$	$4f^0$		187	106
58	Cerium	Ce	$4f^1 5d^1 6s^2$	$4f^2$	$4f^1$	$4f^0$	183	103
59	Praseodymium	Pr	$4f^3 6s^2$	$4f^3$	$4f^2$	$4f^1$	182	101
60	Neodymium	Nd	$4f^4 6s^2$	$4f^4$	$4f^3$	$4f^2$	181	99
61	Promethium	Pm	$4f^5 6s^2$	$4f^5$	$4f^4$		181	98
62	Samarium	Sm	$4f^6 6s^2$	$4f^6$	$4f^5$		180	96
63	Europium	Eu	$4f^7 6s^2$	$4f^7$	$4f^6$		199	95
64	Gadolinium	Gd	$4f^7 5d^1 6s^2$	$4f^7 5d^1$	$4f^7$		180	94
65	Terbium	Tb	$4f^9 6s^2$	$4f^9$	$4f^8$	$4f^7$	178	92
66	Dysprosium	Dy	$4f^{10} 6s^2$	$4f^{10}$	$4f^9$	$4f^8$	177	91
67	Holmium	Ho	$4f^{11} 6s^2$	$4f^{11}$	$4f^{10}$		176	89
68	Erbium	Er	$4f^{12} 6s^2$	$4f^{12}$	$4f^{11}$		175	88
69	Thulium	Tm	$4f^{13} 6s^2$	$4f^{13}$	$4f^{12}$		174	87
70	Ytterbium	Yb	$4f^{14} 6s^2$	$4f^{14}$	$4f^{13}$		173	86
71	Lutetium	Lu	$4f^{14} 5d^1 6s^2$	$4f^{14} 5d^1$	$4f^{14}$	–	–	–

7.5.3 Oxidation States

There is a greater range of oxidation states, which is in part attributed to the fact that the 5*f*, 6*d* and 7*s* levels are of comparable energies. The known oxidation states of actinoids are listed in Table mentioned below.

The actinoids show in general +3 oxidation state. The elements, in the first half of the series frequently exhibit higher oxidation states. For example, the maximum oxidation state increases from +4 in Th to +5, +6 and +7 respectively in Pa, U and Np but decreases in succeeding elements (Table 8.11). The actinoids resemble the lanthanoids in having more compounds in +3 state than in the +4 state. However, +3 and +4 ions tend to hydrolyse. Because the distribution of oxidation states among the actinoids is so uneven and so different for the earlier and latter elements, it is unsatisfactory to review their chemistry in terms of oxidation states.

Table 8.11: Oxidation States of Actinium and Actinoids

Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr
3		3	3	3	3	3	3	3	3	3	3	3	3	3
	4	4	4	4	4	4	4	4						
		5	5	5	5	5								
			6	6	6	6								
				7	7									

General Characteristics and Comparison with Lanthanoids

The actinoid metals are all silvery in appearance but display a variety of structures. The structural variability is obtained due to irregularities in metallic radii which are far greater than in lanthanoids.

The actinoids are highly reactive metals, especially when finely divided. The action of boiling water on them, for example, gives a mixture of oxide and hydride and combination with most non metals takes place at moderate temperatures. Hydrochloric acid attacks all metals but most are slightly affected by nitric acid owing to the formation of protective oxide layers; alkalies have no action.

The magnetic properties of the actinoids are more complex than those of the lanthanoids. Although the variation in the magnetic susceptibility of the actinoids with the number of unpaired 5*f* electrons is roughly parallel to the corresponding results for the lanthanoids, the latter have higher values. It is evident from the behaviour of the actinoids that the ionisation enthalpies of the early actinoids, though not accurately known, but are lower than for the early lanthanoids. This is quite reasonable since it is to be expected that when 5*f* orbitals are beginning to be occupied, they will penetrate less into the inner core of electrons. The 5*f* electrons, will therefore, be more effectively shielded from the nuclear charge than the 4*f* electrons of the corresponding lanthanoids. Because the outer electrons are less firmly held, they are available for bonding in the actinoids A comparison of the actinoids with the lanthanoids, with respect to different characteristics as discussed above, reveals that behaviour similar to that of the lanthanoids is not evident until the second half of the actinoid series. However, even the early actinoids resemble the lanthanoids in showing close similarities with each other and in gradual variation in properties which do not entail change in oxidation state. The lanthanoid and actinoid contractions, have extended effects on the sizes, and therefore, the properties of the elements succeeding them in their respective periods. The lanthanoid contraction is more important because the chemistry of elements succeeding the actinoids is much less known at the present time.

COORDINATION COMPOUNDS

In the previous sections we learnt that the transition metals form a large number of **complex compounds** in which the metal atoms are bound to a number of anions or neutral molecules. In modern terminology such compounds are called **coordination compounds**. The chemistry of coordination compounds is an important and challenging area of modern inorganic chemistry. New concepts of chemical bonding and molecular structure have provided insights into the functioning of vital components of biological systems. Chlorophyll, haemoglobin and vitamin B_{12} are coordination compounds of magnesium, iron and cobalt respectively. Variety of metallurgical processes, industrial catalysts and analytical reagents involve the use of coordination compounds. Coordination compounds also find many applications in electroplating, textile dyeing and medicinal chemistry

7.6 WERNER'S THEORY OF COMPOUNDS

Alfred Werner (1866-1919), a Swiss chemist was the first to formulate his ideas about the structures of coordination compounds. He prepared and characterised a large number of coordination compounds and studied their physical and chemical behaviour by simple experimental techniques. Werner proposed the concept of a **primary valence** and a **secondary valence** for a metal ion. Binary compounds such as $CrCl_3$, $CoCl_2$ or $PdCl_2$ have primary valence of 3, 2 and 2 respectively. In a series of compounds of cobalt(III) chloride with ammonia, it was

found that some of the chloride ions could be precipitated as AgCl on adding excess silver nitrate solution in cold but some remained in solution.

1 mol	$CoCl_3.6NH_3$	(Yellow)	gave	3 mol AgCl
1 mol	$CoCl_3.5NH_3$	(Purple)	gave	2 mol AgCl
1 mol	$CoCl_3.4NH_3$	(Green)	gave	1 mol AgCl
1 mol	$CoCl_3.4NH_3$	(Violet)	gave	1 mol AgCl

These observations, together with the results of conductivity measurements in solution can be explained if (i) six groups in all, either chloride ions or ammonia molecules or both, remain bonded to the cobalt ion during the reaction and (ii) the compounds are formulated as shown in the following Table . where the atoms within the square brackets form a single entity which does not dissociate under the reaction conditions. Werner proposed the term **secondary valence** for the number of groups bound directly to the metal ion; in each of these examples the secondary valences are six.

Colour	Formula	Solution conductivity corresponds to
Yellow	$[Co(NH_3)_6]^{3+}3Cl^-$	1:3 electrolyte
Purple	$[CoCl(NH_3)_5]^{2+}2Cl^-$	1:2 electrolyte
Green	$[CoCl_2(NH_3)_4]^+Cl^-$	1:1 electrolyte
Violet	$[CoCl_2(NH_3)_4]^+Cl^-$	1:1 electrolyte

Note that the last two compounds in Table 9.1 have identical empirical formula, $CoCl_3.4NH_3$, but distinct properties. Such compounds are termed as isomers. Werner in 1898, propounded his theory of coordination compounds. The main postulates are:

In coordination compounds metals show two types of linkages (valences)-primary and secondary

The primary valences are normally ionisable and are satisfied by negative ions.

(1) The secondary valences are non ionisable. These are satisfied by neutral molecules or negative ions. The secondary valence is equal to the coordination number and is fixed for a metal

The ions/groups bound by the secondary linkages to the metal have characteristic spatial arrangements corresponding to different coordination numbers.

In modern formulations, such spatial arrangements are called coordination *polyhedra*. The species within the square bracket are coordination entities or complexes and the ions outside the square bracket are called counter ions.

He further postulated that octahedral, tetrahedral and square planar geometrical shapes are more common in coordination compounds of transition metals. Thus, $[Co(NH_3)_6]^{3+}$, $[CoCl(NH_3)_5]^{2+}$ and $[CoCl_2(NH_3)_4]^+$ are octahedral entities, while $[Ni(CO)_4]$ and $[PtCl_4]^{2-}$ are tetrahedral and square planar, respectively.

7.7 Definitions of some terms

Coordination entity:

A coordination entity constitutes a central metal atom or ion bonded to a fixed number of ions or molecules. For example, $[CoCl_3(NH_3)_3]$ is a coordination entity in which the cobalt ion is

surrounded by three ammonia molecules and three chloride ions. Other examples are $[Ni(CO)_4]$, $[PtCl_2(NH_3)_2]$, $[Fe(CN)_6]^{4-}$, $[Co(NH_3)_6]^{3+.}$

Central atom/ion

In a coordination entity, the atom/ion to which a fixed number of ions/groups are bound in a definite geometrical arrangement around it, is called the central atom or ion. For example, the central atom/ion in the coordination entities: $[NiCl_2(H_2O)_4]$, $[CoCl(NH_3)_5]^{2+}$ and $[Fe(CN)_6]^{3-}$ are Ni^{2+}, Co^{3+} and Fe^{3+}, respectively.

These central atoms/ions are also referred to as **Lewis acids**.

Ligands

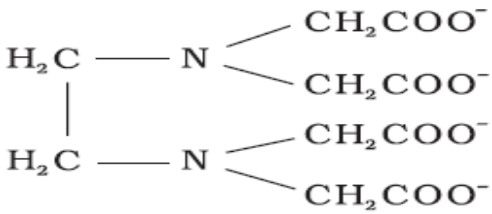
The ions or molecules bound to the central atom/ion in the coordination entity are called ligands. These may be simple ions such as Cl^-, small molecules such as H_2O or NH_3, larger molecules such as $H_2NCH_2CH_2NH_2$ or $N(CH_2CH_2NH_2)_3$ or even macromolecules,

The ions or molecules bound to the central atom/ion in the coordination entity are called ligands. These may be simple ions such as Cl^-, small molecules such as H_2O or NH_3, larger molecules such as $H_2NCH_2CH_2NH_2$ or $N(CH_2CH_2NH_2)_3$ or even macromolecules,

such as proteins. When a ligand is bound to a metal ion through a single donor atom, as with Cl^-, H_2O or NH_3, the ligand is said to be **unidentate.**

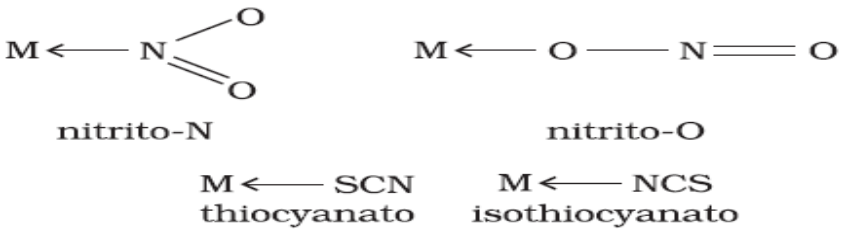
When a ligand can bind through two donor atoms as in $H_2NCH_2CH_2NH_2$ (ethane-1,2-diamine) or $C_2O_4^{2-}$ (oxalate), the ligand is said to be **didentate** and when several donor atoms are present in a single ligand as in $N(CH_2CH_2NH_2)_3$, the ligand is said to be **polydentate**. Ethylenediaminetetraacetate ion ($EDTA^{4-}$) is an important hexadentate ligand. It can bind through two nitrogen and four oxygen atoms to a central metal ion. When a di- or polydentate ligand uses its two or more donor atoms to bind a single metal ion, it is said to be a **chelate** ligand.

The number of such ligating groups is called the **denticity** of the ligand. Such complexes, called chelate complexes tend to be more stable than similar complexes containing unidentate ligands



.

Ligand which can ligate through two different atoms is called **ambidentate ligand**. Examples of such ligands are the NO_2^- and SCN^- ions. NO_2^- ion can coordinate either through nitrogen or through oxygen to a central metal atom/ion. Similarly, SCN^- ion can coordinate through the sulpher or nitrogen atom.



Coordination number

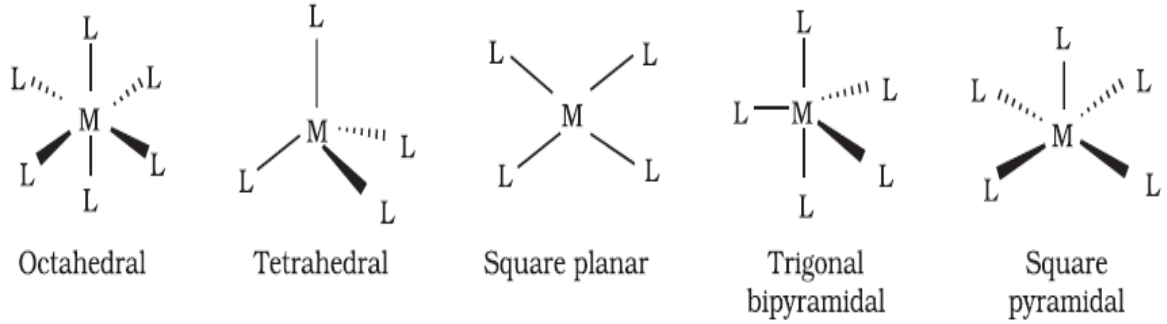
The coordination number (CN) of a metal ion in a complex can be defined as the number of ligand donor atoms to which the metal is directly bonded. For example, in the complex ions, $[PtCl_6]^{2-}$ and $[Ni(NH_3)_4]^{2+}$, the coordination number of Pt and Ni are 6 and 4 respectively. Similarly, in the complex ions, $[Fe(C_2O_4)_3]^{3-}$ and $[Co(en)_3]^{3+}$, the coordination number of both, Fe and Co, is 6 because $C_2O_4^{2-}$ and en (ethane-1,2-diamine) are didentate ligands. It is important to note here that coordination number of the central atom/ion is determined only by the number of sigma bonds formed by the ligand with the central atom/ion. Pi bonds, if formed between the ligand and the central atom/ion, are not counted for this purpose.

Coordination sphere

The central atom/ion and the ligands attached to it are enclosed in square bracket and is collectively termed as the **coordination sphere**. The ionisable groups are written outside the bracket and are called counter ions. For example, in the complex $K_4[Fe(CN)_6]$, the coordination sphere is $[Fe(CN)_6]^{4-}$ and the counter ion is K+.

Coordination polyhedron

The spatial arrangement of the ligand atoms which are directly attached to the central atom/ion defines a coordination polyhedron about the central atom. The most common coordination polyhedra are octahedral, square planar and tetrahedral. For example, $[Co(NH_3)_6]^{3+}$ is octahedral, $[Ni(CO)_4]$ is tetrahedral and $[PtCl_4]^{2-}$ is square planar.



Oxidation number of central atom

The oxidation number of the central atom in a complex is defined as the charge it would carry if all the ligands are removed along with the electron pairs that are shared with the central atom. The oxidation number is represented by a Roman numeral in parenthesis following the name of the coordination entity. For example, oxidation number of copper in $[Cu(CN)_4]^{3-}$ is +1 and it is written as Cu(I).

Homoleptic and heteroleptic complexes

Complexes in which a metal is bound to only one kind of donor groups, *e.g.*, $[Co(NH_3)_6]^{3+}$, are known as homoleptic. Complexes in which a metal is bound to more than one kind of donor groups,

e.g., $[Co(NH_3)_4Cl_2]^{+}$, are known as heteroleptic.

7.8 NOMENCLATURE

Nomenclature is important in Coordination Chemistry because of the need to have an unambiguous method of describing formulas and writing systematic names, particularly when dealing with isomers. The formulas and names adopted for coordination entities are based on the recommendations of the International Union of Pure and Applied Chemistry (IUPAC).

The names of coordination compounds are derived by following the principles of additive nomenclature. Thus, the groups that surround the central atom must be identified in the name. They are listed as prefixes to the name of the central atom along with any appropriate multipliers.

The following rules are used when naming coordination compounds:

(i) The cation is named first in both positively and negatively charged coordination entities.

(ii) The ligands are named in an alphabetical order before the name of the central atom/ion. (This procedure is reversed from writing formula).

(iii) Names of the anionic ligands end in –o, those of neutral and cationic ligands are the same except aqua for H_2O, ammine for NH_3, carbonyl for CO and nitrosyl for NO. These are placed with in enclosing marks ( ).

(iv) Prefixes mono, di, tri, etc., are used to indicate the number of the individual ligands in the coordination entity. When the names of the ligands include a numerical prefix, then the terms, *bis*, *tris*, *tetrakis* are used, the ligand to which they refer being placed in parentheses. For example, $[NiCl_2(PPh_3)_2]$ is named as dichlorobis(triphenylphosphine)nickel(II).

(v) Oxidation state of the metal in cation, anion or neutral coordination entity is indicated by Roman numeral in parenthesis.

If the complex ion is a cation, the metal is named same as the element. For example, Co in a complex cation is called cobalt and Pt is called platinum. If the complex ion is an anion, the name of the metal ends with the suffix – ate. For example, Co in a complex anion $[Co(SCN)_4]^{2-}$ is called cobaltate. For some metals, the Latin names are used in the complex anions, *e*.g., ferrate for Fe. names are used in the complex anions, *e*.g., ferrate for Fe.

(vi) The neutral complex molecule is named similar to that of the complex cation.

7.9 ISOMERISM IN COMPLEX COMPOUNDS

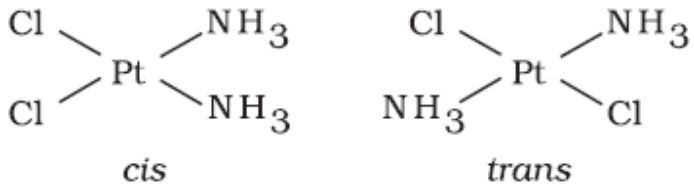
Isomers are two or more compounds that have the same chemical formula but a different arrangement of atoms. Because of the different arrangement of atoms, they differ in one or more physical or chemical properties. Two principal types of isomerism are known among coordination compounds. Each of which can be further subdivided.

(a) Stereoisomerism
- (i) Geometrical isomerism
- (ii) Optical isomerism

(b) Structural isomerism
- (i) Linkage isomerism
- (ii) Coordination isomerism
- (iii) Ionisation isomerism
- (iv) Solvate isomerism

Stereoisomers have the same chemical formula and chemical bonds but they have different spatial arrangement. Structural isomers have different bonds. A detailed account of these isomers are given below.

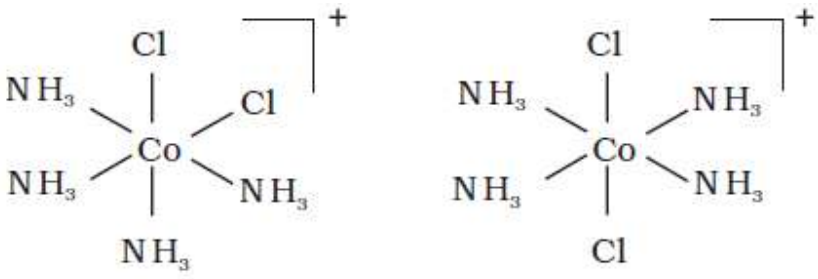
This type of isomerism arises in heteroleptic complexes due to different possible geometric arrangements of the ligands. Important examples of this behaviour are found with coordination numbers 4 and 6. In a square planar complex of formula $[MX_2L_2]$ (X and L are unidentate), the two ligands X may be arranged adjacent to each other in a ***cis*** isomer, or opposite to each other in a ***trans*** isomer



cis *trans*

Geometrical isomers (cis and trans) of $Pt(NH_3)_2Cl_2$)

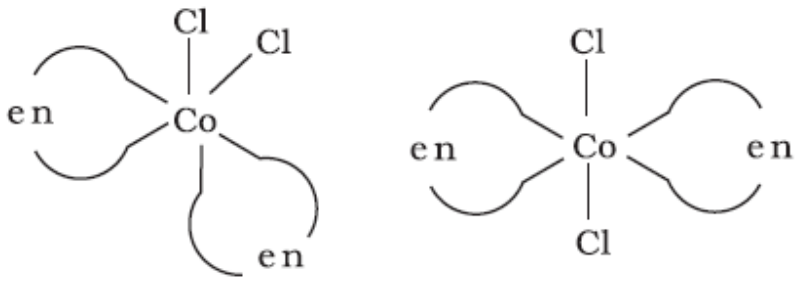
Other square planar complex of the type MABXL (where A, B, X, L are unidentates) shows three isomers-two *cis* and one *trans*. You may attempt to draw these structures. Such isomerism is not possible for a tetrahedral geometry but similar behaviour is possible in octahedral complexes of formula $[MX_2L_4]$ in which the two ligands X may be oriented ***cis***or ***trans*** to each other



cis *trans*

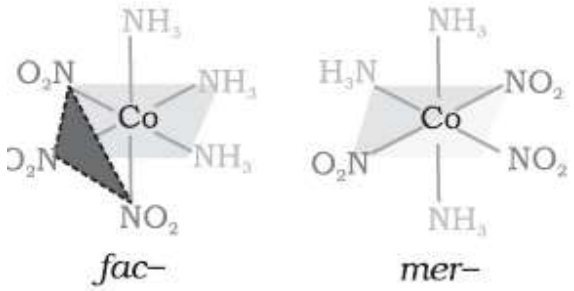
Geometrical isomers (cis and trans) of $[Co(NH_3)_4Cl_2]^+$

This type of isomerism also arises when didentate ligands L – L [*e.g.*, $NH_2\ CH_2\ CH_2\ NH_2$ (en)] are present in complexes of formula$[MX_2(L – L)2]$



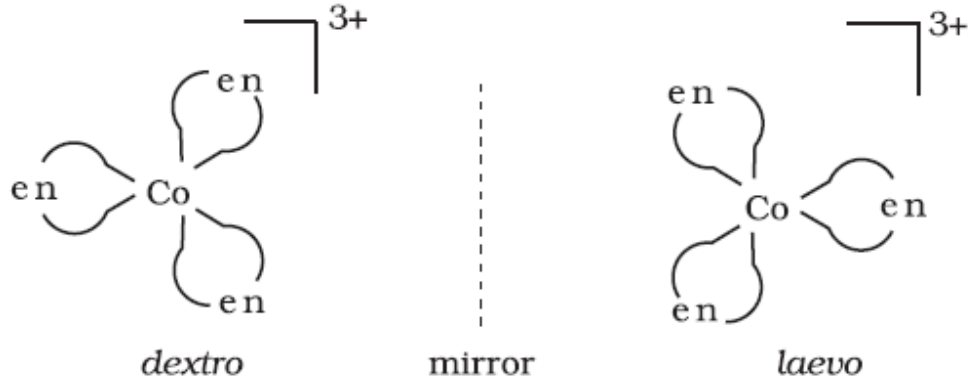
Geometrical isomers (cis and trans) of $[CoCl_2(en)_2]$

Another type of geometrical isomerism occurs in octahedral coordination entities of the type $[Ma_3b_3]$ like $[Co(NH_3)_3(NO_2)_3]$. If three donor atoms of the same ligands occupy adjacent positions at the corners of an octahedral face, we have the **facial (fac) isomer**. When the positions are around the meridian of the octahedron, we get the **meridional(mer) isomer**



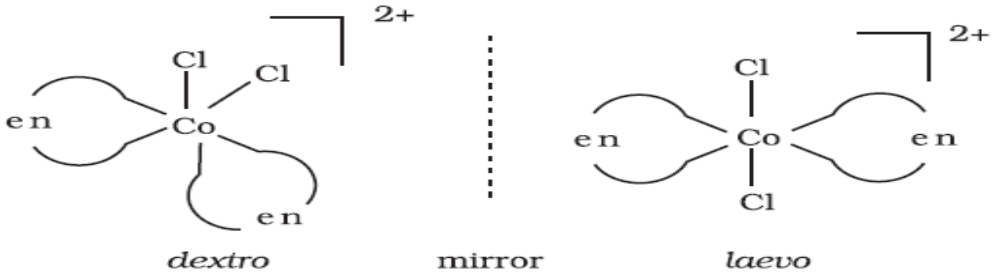
Optical Isomerism

Optical isomers are mirror images that cannot be superimposed on one another. These are called as *enantiomers*. The molecules or ions that cannot be superimposed are called *chiral*. The two forms are called *dextro (d)* and *laevo (l)* depending upon the direction they rotate the plane of polarised light in a polarimeter (*d* rotates to the right, *l* to the left). Optical isomerism is common in octahedral complexes involvingdidentate ligands.



Optical isomers (d and l) of $[Co(en)_3]^{3+}$

In a coordination entity of the type $[PtCl2(en)_2]^{2+}$, only the *cis*-isomer shows opticalactivity



Optical isomer (d and l) of cis-$[PtCl_2(en)_2]^{2+}$

Linkage Isomerism

Linkage isomerism arises in a coordination compound containing ambidentate ligand. A simple example is provided by complexes containing the thiocyanate ligand, NCS^-, which may bind through the nitrogen to give M–NCS or through sulphur to give M–SCN. Jørgensen discovered such behaviour in the complex $[Co(NH_3)_5(NO_2)]Cl_2$, which is obtained as the red form, in which the nitrite ligand is bound through oxygen (–ONO), and as the yellow form, in which the nitrite ligand is bound through nitrogen ($–NO_2$).

Coordination Isomerism

This type of isomerism arises from the interchange of ligands between cationic and anionic entities of different metal ions present in a complex. An example is provided by $[Co(NH_3)_6][Cr(CN)_6]$, in which the NH_3 ligands are bound to Co^{3+} and the CN^- ligands to Cr^{3+}. In its coordination isomer $[Cr(NH_3)_6][Co(CN)_6]$, the NH_3 ligands are bound to Cr^{3+} and theCN^- ligands to Co^{3+}

Ionisation Isomerism

This form of isomerism arises when the counter ion in a complex salt is itself a potential ligand and can displace a ligand which can then become the counter ion. An example is provided by the ionisation isomers $[Co(NH_3)_5SO_4]Br$ and $[Co(NH3)_5Br]SO_4$

Solvate Isomerism

This form of isomerism is known as 'hydrate isomerism' in case where water is involved as a solvent. This is similar to ionisation isomerism. Solvate isomers differ by whether or not a solvent molecule is directly bonded to the metal ion or merely present as free solvent molecules in the crystal lattice. An example is provided by the aqua complex $[Cr(H_2O)_6]Cl_3$ (violet) and its solvate isomer $[Cr(H_2O)_5Cl]Cl_2.H_2O$(grey-green)

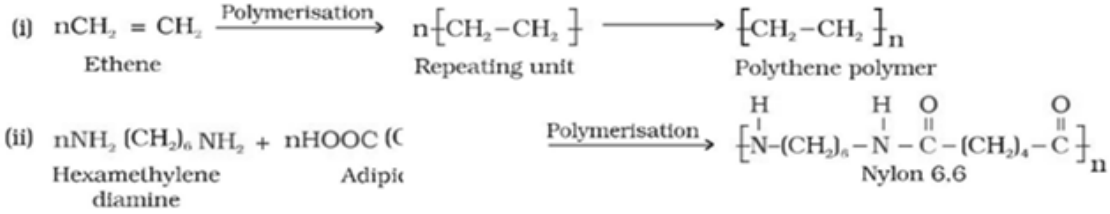
IMPORTANT QUESTIONS:

1. What are transition elements ? Give examples.
2. Why are d-block elements called transition elements?
3. Write the electronic configuration of chromium (Cr) and copper (Cu).
4. Scandium is a transition element.But Zn is not.Why ?
5. What is an alloy ? Give example.
6. What is lanthanoid contraction.
7. What is a ligand?
8. Write the charecteristics of interstitial compounds.
9. Explain Werner's theory of coordination compounds with suitable examples.
10. What is misch metal? Give its composition and uses.

CHAPTER 8

POLYMERS

The word 'polymer' is coined from two Greek words: *poly* means many and *mer* means unit or part. The term polymer is defined as very large molecules having high molecular mass (10^3-10^7u). These are also referred to as **macromolecules,** which are formed by joining of repeating structural units on a large scale. The repeating structural units are derived from some simple and reactive molecules known as monomers and are linked to each other by covalent bonds. This process of formation of polymers from respective monomers is called **polymerisation**. The transformation of ethene to polythene and interaction of hexamethylene diamine and adipic acid leading to the formation of Nylon 6, 6 are examples of two different types of polymerisation reactions.



8.1 Classification of Polymers

There are several ways of classification of polymers based on some special considerations. The following are some of the common classifications of polymers:

8.1.1 Classification Based on Source

Under this type of classification, there are three sub categories.

1. ***Natural polymers***

   These polymers are found in plants and animals. Examples are proteins, cellulose, starch, resins and rubber.

2. ***Semi-synthetic polymers***

   Cellulose derivatives as cellulose acetate (rayon) and cellulose nitrate, etc. are the usual examples of this sub category.

3. ***Synthetic polymers***

   A variety of synthetic polymers as plastic (polythene), synthetic fibres (nylon 6,6) and synthetic rubbers (Buna - S) are examples of man-made polymers extensively used in daily life as well as in industry.

8.1.2 Classification Based on Source

Under this type of classification, there are three sub categories.

1. ***Natural polymers***

   These polymers are found in plants and animals. Examples are proteins, cellulose, starch, resins and rubber.

2. ***Semi-synthetic polymers***

   Cellulose derivatives as cellulose acetate (rayon) and cellulose nitrate, etc. are the usual examples of this sub category.

3. ***Synthetic polymers***A variety of synthetic polymers as plastic (polythene), synthetic fibres (nylon 6,6) and synthetic rubbers (Buna - S) are examples of man-made polymers extensively used in daily life as well as in industry.

8.1.3 Classification Based on Structure of Polymers

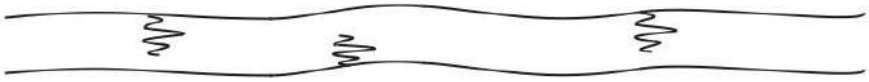
There are three different types based on the structure of the polymers.

1. Linear polymers

These polymers consist of long and straight chains. The examples are high density polythene, polyvinyl chloride, etc. These are represented as:

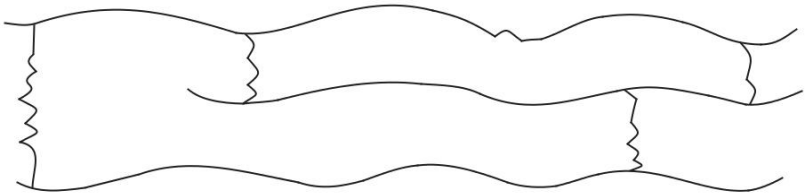
2. Branched chain polymers

These polymers contain linear chains having some branches, *e.g.*, low density polythene. These are depicted as follows:



3. Cross linked or Network polymers

These are usually formed from bi-functional and tri-functional monomers and contain strong covalent bonds between various linear polymer chains, e.g. bakelite, melamine, etc. These polymers are depicted as follows:



8.1.4 Classification Based on Mode of Polymerisation

Polymers can also be classified on the basis of mode of polymerisation into two sub groups.

1. Addition polymers

The addition polymers are formed by the repeated addition of monomer molecules possessing double or triple bonds, *e.g.*, the formation of polythene from ethene and polypropene from propene. However, the addition polymers formed by the polymerisation of a single monomeric species are known as **homopolymers**, *e.g.*, polythene.

$$\underset{\text{Ethene}}{n\,CH_2=CH_2} \longrightarrow \underset{\text{Polythene}}{-(CH_2-CH_2)-_n} \quad \text{Homopolymer}$$

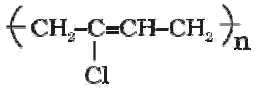
The polymers made by addition polymerisation from two different monomers are termed as **copolymers**, *e.g.*, Buna-S, Buna-N, etc.

$$\underset{\text{1, 3-Butadien}}{n\,CH_2=CH-CH=CH_2} + \underset{\text{Styrene}}{n\,C_6H_5CH=CH_2} \longrightarrow \underset{\substack{\text{Butadiene-styrene copolymer}\\ \text{(Buna - S)}}}{-(CH_2-CH=CH-CH_2-CH_2-\underset{C_6H_5}{CH})_n-}$$

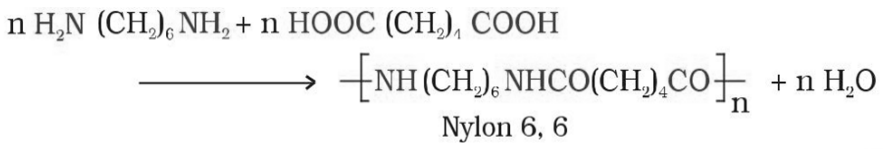
2. Condensation polymers

2. Condensation polymers

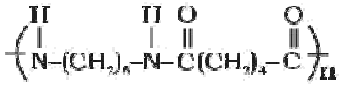
The condensation polymers are formed by repeated condensation reaction between two different bi-functional or tri-functional monomeric units. In these polymerisation reactions, the elimination of small molecules such as water, alcohol, hydrogen chloride, etc. take place. The examples are terylene (dacron), nylon 6, 6, nylon 6, etc. For example, nylon 6, 6 is formed by the condensation of hexamethylene diamine with adipic acid.



Neoprene



Nylon 6, 6



Nylon 6,6

8.1.5 Classification Based on Molecular Forces

A large number of polymer applications in different fields depend on their unique mechanical properties like tensile strength, elasticity, toughness, etc. These mechanical properties are governed by intermolecular forces, *e.g.*, van der Waals forces and hydrogen bonds, present in the polymer. These forces also bind the polymer chains. Under this category, the polymers are classified into the following four sub groups on the basis of magnitude of intermolecular forces present in them.

1. Elastomers

These are rubber – like solids with elastic properties. In these elastomeric polymers, the polymer chains are held together by the weakest intermolecular forces. These weak binding forces permit the polymer to be stretched. A few 'crosslinks' are introduced in between the chains, which help the polymer to retract to its original position after the force is released as in vulcanised rubber. The examples are buna-S, buna-N, neoprene, etc.

2. Fibres

Fibres are the thread solids which possess high tensile strength and high modulus. These characteristics can forming be attributed to the strong intermolecular forces like hydrogen bonding. These strong forces also lead to close packing of chains and thus impart crystalline nature. The examples are polyamides (nylon 6, 6), polyesters (terylene), etc.

3. Thermoplastic polymers

These are the linear or slightly branched long chain molecules capable of repeatedly softening on heating and hardening on cooling. These polymers possess intermolecular forces of attraction intermediate between elastomers and fibres. Some common thermoplastics are polythene, polystyrene, polyvinyls, etc.

4.Thermosetting polymers

These polymers are cross linked or heavily branched molecules, which on heating undergo extensive cross linking in moulds and again become infusible. These cannot be reused. Some common examples are bakelite, urea-formaldelyde resins, etc.

8.1.6 Classification Based on Growth Polymerisation

The addition and condensation polymers are nowadays also referred as chain growth polymers and step growth polymers depending on the type of polymerisation mechanism they undergo during their formation.

8.2 Types of Polymerisation Reactions

There are two broad types of polymerisation reactions, *i.e.*, the addition or chain growth polymerisation and condensation or step growth polymerisation.

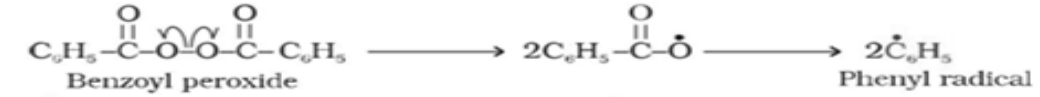
8.2.1 Addition Polymerisation or Chain Growth Polymerisation

In this type of polymerisation, the molecules of the same monomer or diferent monomers add together on a large scale to form a polymer. The monomers used are unsaturated compounds, *e.g.*, alkenes, alkadienes and their derivatives. This mode of polymerisation leading to an increase in chain length or chain growth can take place through the formation of either free radicals or ionic species. However, the free radical governed addition or chain growth polymerisation is the most common mode.

1. Free radical mechanism

A variety of alkenes or dienes and their derivatives are polymerised in the presence of a free radical generating initiator (catalyst) like benzoyl peroxide, acetyl peroxide, tert-butyl peroxide, etc. For example, the polymerisation of ethene to polythene consists of heating or exposing to light a mixture of ethene with a small amount of benzoyl peroxide initiator. The process starts with the addition of phenyl free radical formed by the peroxide to the ethene double bond thus generating a new and larger free radical. This step is called **chain initiating step**. As this radical reacts with another molecule of ethene, another bigger sized radical is formed. The repetition of this sequence with new and bigger radicals carries the reaction forward and the step is termed as **chain propagating step**. Ultimately, at some stage the product radical thus formed reacts with another radical to form the polymerised product. This step is called the **chain terminating step**. The sequence of steps may be depicted as follows:

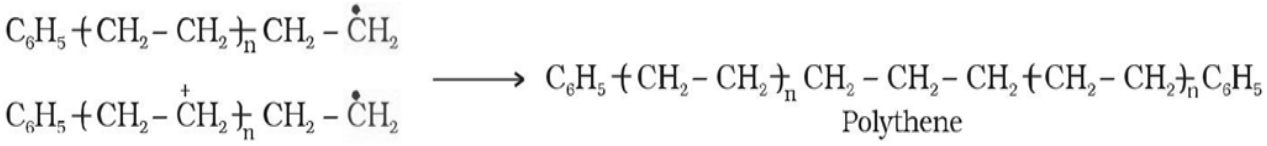
Chain initiation steps



Chain propagating step

Chain terminating step

For termination of the long chain, these free radicals can combine in different ways to form polythene. One mode of termination of chain is shown as under:



2 Preparation of some important addition polymers

(a) Polythene

There are two types of polythene as given below:

(i) *Low density polythene*:

It is obtained by the polymerisation of ethene under high pressure of 1000 to 2000 atmospheres at a temperature of 350 K to 570 K in the presence of traces of dioxygen or a peroxide initiator (catalyst). The low density polythene (LDP) obtained through the free radical addition and H-atom abstraction has highly branched structure. Low density polythene is chemically inert and tough but flexible and a poor conductor of electricity. Hence, it is used in the insulation of electricity carrying wires and manufacture of squeeze bottles, toys and flexible pipes.

(ii) *High density polythene:*

It is formed when addition polymerisation of ethene takes place in a hydrocarbon solvent in the presence of a catalyst such as triethylaluminium and titanium tetrachloride (Ziegler-Natta catalyst) at a temperature of 333 K to 343 K and under a pressure of 6-7 atmospheres. High density polythene (HDP) thus produced, consists of linear molecules and has a high density due to close packing. It is also chemically inert and tougher and harder. It is used for manufacturing buckets, dustbins, bottles, pipes, etc.

(b) Polytetrafluoroethene (Teflon)

Teflon is manufactured by heating tetrafluoroethene with a free radical or persulphate catalyst at high pressures. It is chemically inert and resistant to attack by corrosive reagents. It is used in making oil seals and gaskets and also used for non – stick surface coated utensils.

$$\underset{\text{Tetrafluoroethene}}{n\ CF_2 = CF_2} \xrightarrow[\text{High pressure}]{\text{Catalyst}} \underset{\text{Teflon}}{\left[CF_2 - CF_2\right]_n}$$

(c) Polyacrylonitrile

The addition polymerisation of acrylonitrile in presence of a peroxide catalyst leads to the formation of polyacrylonitrile.

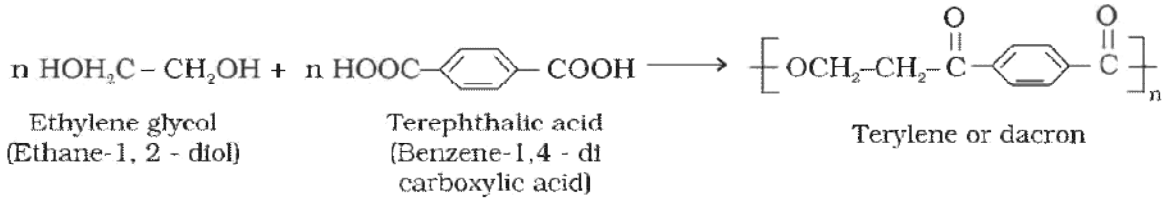
$$\underset{\text{Acrylonitrile}}{n\ CH_2 = CHCN} \xrightarrow[\text{Peroxide catalyst}]{\text{Polymerisation}} \underset{\text{Polyacrylonitrile}}{\left[CH_2 - \underset{\displaystyle CN}{\underset{|}{C}}H\right]_n}$$

Polyacrylonitrile is used as a substitute for wool in making commercial fibres as orlon or acrilan.

8.2.2 Condensation Polymerization

This type of polymerization generally involves a repetitive condensation reaction between two bi-functional monomers. These polycondensation reactions may result in the loss of some simple molecules as water, alcohol, etc., and lead to the formation of high molecular mass condensation polymers.

In these reactions, the product of each step is again a bi-functional species and the sequence of condensation goes on. Since, each step produces a distinct functionalised species and is independent of each other, this process is also called as step growth polymerisation.



Ethylene glycol (Ethane-1, 2 - diol) | Terephthalic acid (Benzene-1,4 - di carboxylic acid) | Terylene or dacron

The formation of terylene or dacron by the interaction of ethylene glycol and terephthalic acid is an example of this type of polymerisation.

Ethylene glycol (Ethane-1, 2 - diol) | Terephthalic acid (Benzene-1,4 - di carboxylic acid) | Terylene or dacron

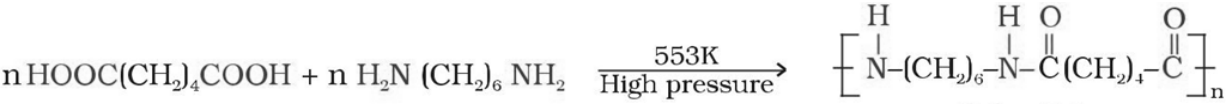
Some important condensation polymerisation reactions characterised by their linking units are described below:

1. Polyamides

These polymers possessing amide linkages are important examples of synthetic fibres and are termed as nylons. The general method of preparation consists of the condensation polymerisation of diamines with dicarboxylic acids and also of amino acids and their lactams.

(a) Preparation of nylons

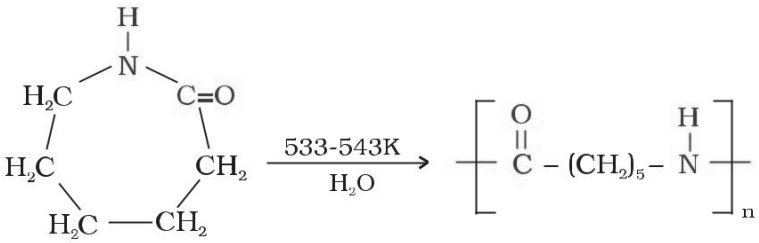
(i) ***Nylon 6,6:*** It is prepared by the condensation polymerization of hexamethylenediamine with adipic acid under high pressure and at high temperature.



Nylon 6,6

Nylon 6, 6 is used in making sheets, bristles for brushes and in textile industry.

(ii) ***Nylon 6:*** It is obtained by heating caprolactum with water at a high temperature.



Caprolactam | Nylon 6

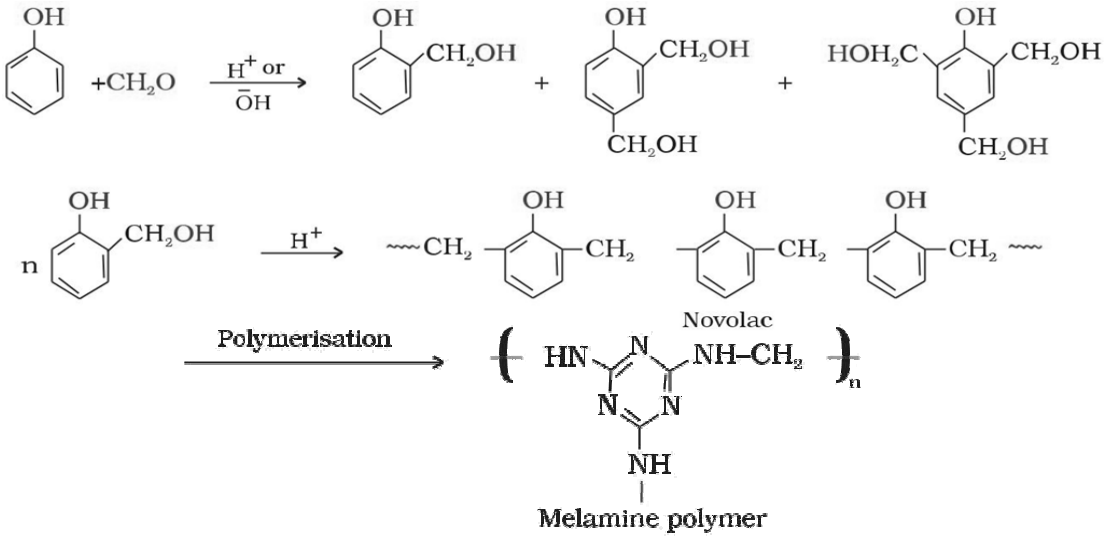
Nylon 6 is used for the manufacture of tyre cords, fabrics and ropes.

2. Polyesters

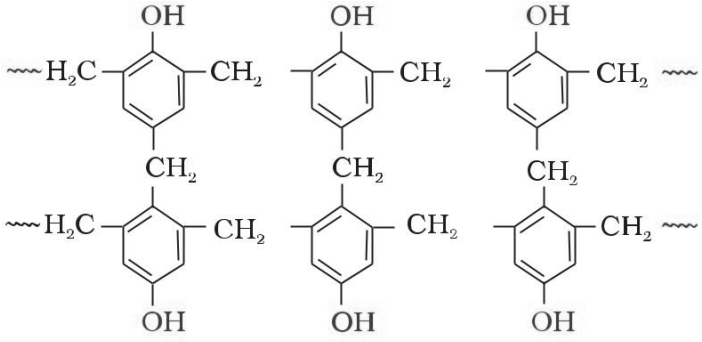
These are the polycondensation products of dicarboxylic acids and diols. Dacron or terylene is the best known example of polyesters. It is manufactured by heating a mixture of ethylene glycol and terephthalic acid at 420 to 460 K in the presence of zinc acetate-antimony trioxide catalyst as per the reaction given earlier. Dacron fibre (terylene) is crease resistant and is used in blending with cotton and wool fibres and also as glass reinforcing materials in safety helmets, etc.

3. Phenol - formaldehyde polymer (Bakelite and related polymers)

Phenol - formaldehyde polymers are the oldest synthetic polymers. These are obtained by the condensation reaction of phenol with formaldehyde in the presence of either an acid or a base catalyst. The reaction starts with the initial formation of *o*-and/or *p*-hydroxymethylphenol derivatives, which further react with phenol to form compounds having rings joined to each other through $-CH_2$ groups. The initial product could be a linear product – **Novolac** used in paints.



Novolac on heating with formaldehyde undergoes cross linking to form an infusible solid mass called **bakelite**. It is used for making combs, phonograph records, electrical switches and handles of various utensils.



Bakelite

4. Melamine – formaldehyde polymer

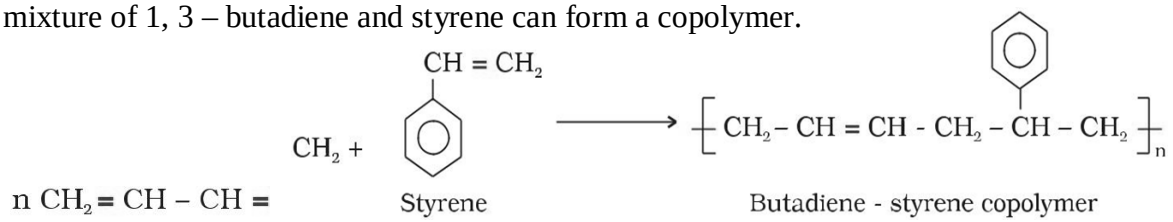
Melamine formaldehyde polymer is formed by the condensation polymerisation of melamine and formaldehyde.

Melamine Formaldehyde Resin intermediate

It is used in the manufacture of unbreakable crockery.

8.2.3 Co-polymerisation

Copolymerisation is a polymerisation reaction in which a mixture of more than one monomeric species is allowed to polymerise and form a copolymer. The copolymer can be made not only by chain growth polymerisation but by step growth polymerisation also. It contains multiple units of each monomer used in the same polymeric chain. For example, a mixture of 1, 3 – butadiene and styrene can form a copolymer.



n $CH_2 = CH - CH =$ Styrene Butadiene - styrene copolymer

1, 3-Butadiene

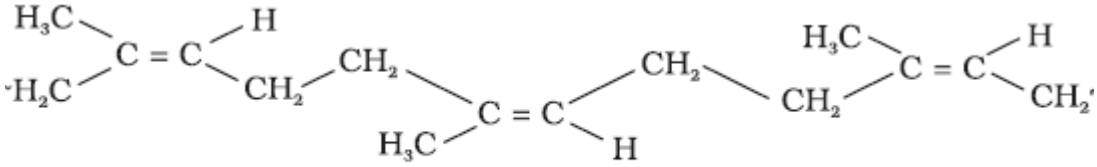
Copolymers have properties quite different from homopolymers. For example, butadiene - styrene copolymer is quite tough and is a good substitute for natural rubber. It is used for the manufacture of autotyres, floortiles, footwear components, cable insulation, etc.

8.2.4 Rubber

1. Natural rubber

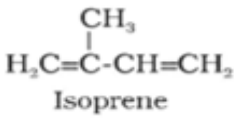
Rubber is a natural polymer and possesses elastic properties. It is also termed as elastomer and has a variety of uses. It is manufactured from rubber latex which is a colloidal dispersion of rubber in water. This latex is obtained from the bark of rubber tree and is found in India, Srilanka, Indonesia, Malaysia and South America.

Natural rubber may be considered as a linear polymer of isoprene (2-methyl-1, 3-butadiene) and is also called as *cis* - 1, 4 - polyisoprene.



Natural rubber

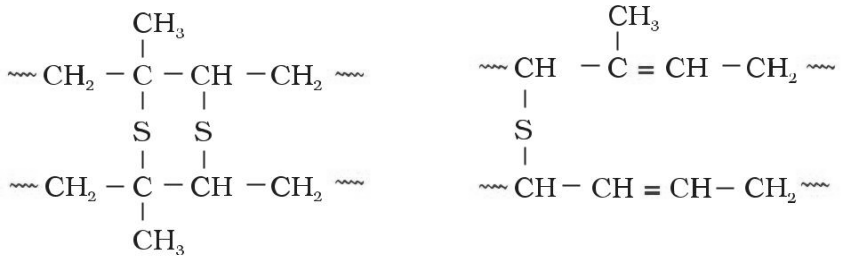
The *cis*-polyisoprene molecule consists of various chains held together by weak van der Waals interactions and has a coiled structure. Thus, it can be stretched like a spring and exhibits elastic properties.



Isoprene

Vulcanisation of rubber: Natural rubber becomes soft at high temperature (>335 K) and brittle at low temperatures (<283 K) and shows high water absorption capacity. It is soluble in non-polar solvents and is non-resistant to attack by oxidising agents. To improve upon these physical properties, a process of vulcanisation is carried out. This process consists of heating a mixture of raw rubber with sulphur and an appropriate additive at a temperature range between 373 K to 415 K. On vulcanisation, sulphur forms cross links at the reactive sites of double bonds and thus the rubber gets stiffened.

In the manufacture of tyre rubber, 5% of sulphur is used as a crosslinking agent. The probable structures of vulcanised rubber molecules are depicted below:



2. Synthetic rubbers

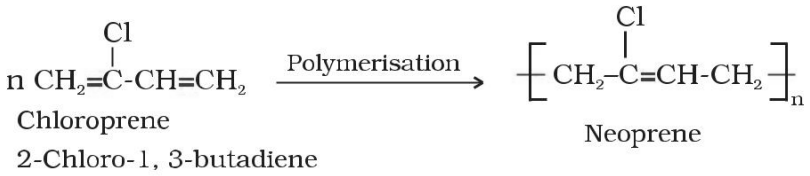
Synthetic rubber is any vulcanisable rubber like polymer, which is capable of getting stretched to twice its length. However, it returns to its original shape and size as soon as the external stretching force is released. Thus, synthetic rubbers are either homopolymers of 1, 3

- butadiene derivatives or copolymers of 1, 3 - butadiene or its derivatives with another unsaturated monomer.

Preparation of Synthetic Rubbers

1. Neoprene

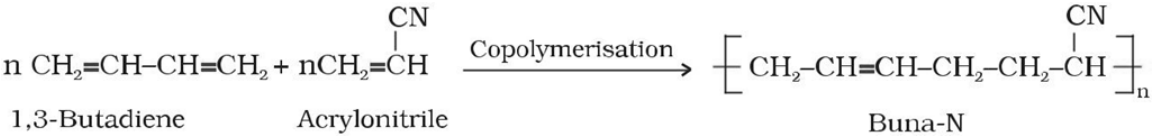
Neoprene or polychloroprene is formed by the free radical polymerisation of chloroprene.



It has superior resistance to vegetable and mineral oils. It is used for manufacturing conveyor belts, gaskets and hoses.

2. Buna – N

You have already studied about Buna-S, in Section 15.1.3. Buna –N is obtained by the copolymerisation of 1, 3 – butadiene and acrylonitrile in the presence of a peroxide catalyst.



It is resistant to the action of petrol, lubricating oil and organic solvents. It is used in making oil seals, tank lining, etc.

8. 3 Molecular Mass of Polymers

Polymer properties are closely related to their molecular mass, size and structure. The growth of the polymer chain during their synthesis is dependent upon the availability of the monomers in the reaction mixture. Thus, the polymer sample contains chains of varying lengths and hence its molecular mass is always expressed as an average. The molecular mass of polymers can be determined by chemical and physical methods.

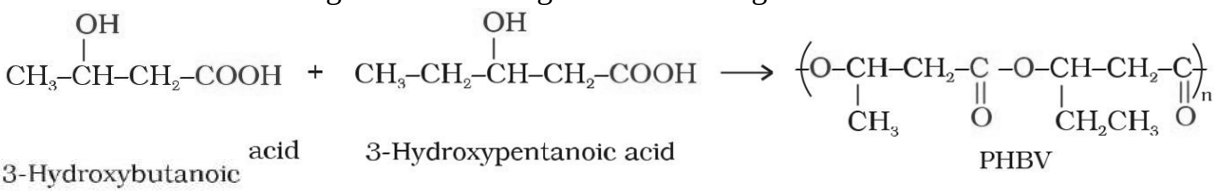
8.4 Biodegradable Polymers

A large number of polymers are quite resistant to the environmental degradation processes and are thus responsible for the accumulation of polymeric solid waste materials. These solid wastes cause acute environmental problems and remain undegraded for quite a long time. In view of the general awareness and concern for the problems created by the polymeric solid wastes, certain new biodegradable synthetic polymers have been designed and developed. These polymers contain functional groups similar to the functional groups present in biopolymers.

Aliphatic polyesters are one of the important classes of biodegradable polymers. Some important examples are given below:

1. Poly β-hydroxybutyrate – co-β-hydroxy valerate (PHBV)

It is obtained by the copolymerisation of 3-hydroxybutanoic acid and 3 - hydroxypentanoic acid. PHBV is used in speciality packaging, orthopaedic devices and in controlled release of drugs. PHBV undergoes bacterial degradation in the environment.



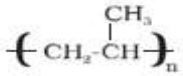
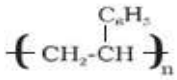
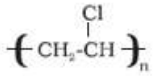
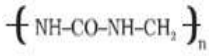
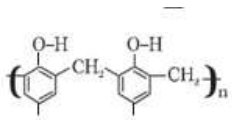
2. Nylon 2–nylon 6

It is an alternating polyamide copolymer of glycine ($H_2N–CH_2–COOH$) and amino caproic acid [$H_2N\ (CH_2)_5\ COOH$] and is biodegradable. Can you write the structure of this copolymer?

8. 5 Polymers of Commercial Importance

Besides, the polymers already discussed, some other commercially important polymers along with their structures and uses are given below in Table 8.1.

Table 8.1: Some Other Commercially Important Polymers

Name of Polymer	Monomer	Structure	Uses
Polypropene	Propene		Manufacture of ropes, toys, pipes, fibres, etc.
Polystyrene	Styrene		As insulator, wrapping material, manufacture of toys, radio and television cabinets.
Polyvinyl chloride (PVC)	Vinyl chloride		Manufacture of rain coats, hand bags, vinyl flooring, water pipes.
Urea-formaldehyle Resin	(a) Urea (b) Formaldehyde		For making unbreakable cups and laminated sheets.
Glyptal	(a) Ethylene glycol (b) Phthalic acid		Manufacture of paints and lacquers.
Bakelite	(a) Phenol (b) Formaldehyde		For making combs, electrical switches, handles of utensils and computer discs.

Summary

Polymers are defined as high molecular mass **macromolecules**, which consist of repeating structural units derived from the corresponding **monomers**. These polymers may be of natural or synthetic origin and are classified in a number of ways. In the presence of an organic peroxide initiator, the alkenes and their derivatives undergo **addition polymerisation** or **chain growth polymerisation** through a **free radical mechanism**. Polythene, teflon, orlon, etc. are formed by addition polymerisation of an appropriate alkene or its derivative. **Condensation polymerisation** reactions are shown by the interaction of bi – or poly functional monomers containing – NH_2, – OH and – COOH groups. This type of polymerisation proceeds through the elimination of certain simple molecules as H_2O, CH_3OH, etc. Formaldehyde reacts with phenol and melamine to form the corresponding condensation polymer products. The condensation polymerisation progresses through step by step and is also called as **step growth polymerisation**. Nylon, bakelite and dacron are some of the important examples of condensation polymers. However, a mixture of two unsaturated monomers exhibits **copolymerisation** and forms a **co-polymer** containing multiple units of each monomer. Natural rubber is a *cis* 1, 4-polyisoprene and can be made more tough by the

process of **vulcanisation** with sulphur. Synthetic rubbers are usually obtained by copolymerisation of alkene and 1, 3 butadiene derivatives.

In view of the potential environmental hazards of synthetic polymeric wastes, certain **biodegradable polymers** such as PHBV and Nylon-2- Nylon-6 are developed as alternatives.

IMPORTANT QUESTIONS

1. Distinguish between the terms Homo Polymer and Copolymer. Give one example of each
2. Define Thermoplastics and Thermosetting polymers with two examples of each
3. Write the names of the monomers for the following polymers
   (i) Polyvinyl Chloride
   (ii) Teflon
   (iii) Bakelite
   (iv) Polystyrene
   (v) BUNA – S
   (vi) BUNA – N
4. What is PHBV? How is it useful to man?
5. What is Ziegler – Natta Catalyst

CHAPTER 9

Biomolecules

A living system grows, sustains and reproduces itself. The most amazing thing about a living system is that it is composed of non-living atoms and molecules. The pursuit of knowledge of what goes on chemically within a living system falls in the domain of ***biochemistry***. Living systems are made up of various complex biomolecules like carbohydrates, proteins, nucleic acids, lipids, etc. Proteins and carbohydrates are essential constituents of our food. These biomolecules interact with each other and constitute the molecular logic of life processes. In addition, some simple molecules like vitamins and mineral salts also play an important role in the functions of organisms. Structures and functions of some of these biomolecules are discussed in this Unit.

9.1 Carbohydrates

Carbohydrates are primarily produced by plants and form a very large group of naturally occurring organic compounds. Some common examples are cane sugar, glucose, starch, etc. Most of them have a general formula, $Cx(H_2O)y$, and were considered as hydrates of carbon from where the name carbohydrate was derived. For example, the molecular formula of glucose ($C_6H_{12}O_6$) fits into this general formula, $C_6(H_2O)_6$. But all the compounds which fit into this formula may not be classified as carbohydrates. Acetic acid (CH_3COOH) fits into this general formula, $C_2(H_2O)_2$ but is not a carbohydrate. Similarly, rhamnose, $C_6H_{12}O_5$ is a carbohydrate but does not fit in this definition. A large number of their reactions have shown that they contain specific functional groups. Chemically, *the carbohydrates may be defined as optically active polyhydroxy aldehydes or ketones or the compounds which produce such units on hydrolysis.* Some of the carbohydrates.

9.1.1 Classification of Carbohydrates

Which are sweet in taste, are also called sugars. The most common sugar, used in our homes is named as sucrose whereas the sugar present in milk is known as lactose. Carbohydrates are also called saccharides (Greek: *sakcharon* means sugar).

Carbohydrates are classified on the basis of their behaviour on hydrolysis. They have been broadly divided into following three groups.

(i) *Monosaccharides***:** A carbohydrate that cannot be hydrolysed further to give simpler unit of polyhydroxy aldehyde or ketone is called a monosaccharide.About 20 monosaccharides are known to occur in nature. Some common examples are glucose, fructose, ribose, etc.

(ii) *Oligosaccharides***:** Carbohydrates that yield two to ten monosaccharide units, on hydrolysis, are called oligosaccharides. They are further classified as disaccharides, trisaccharides, tetrasaccharides, etc., depending upon the number of monosaccharides, they provide on hydrolysis. Amongst these the most common are disaccharides. The two monosaccharide units obtained on hydrolysis of a disaccharide may be same or different. For example, sucrose on hydrolysis gives one molecule each of glucose and fructose whereas maltose gives two molecules of glucose only.

(iii) *Polysaccharides***:** Carbohydrates which yield a large number of monosaccharide units on hydrolysis are called polysaccharides.Some common examples are starch, cellulose, glycogen, gums, etc. Polysaccharides are not sweet in taste, hence they are also called non-sugars. The carbohydrates may also be classified as either reducing or nonreducing sugars. All those carbohydrates which reduce Fehling's solution and Tollens' reagent are referred to as reducing sugars. All monosaccharides whether aldose or ketose are *reducing sugars.* In disaccharides, if the reducing groups of monosaccharides i.e., aldehydic or ketonic groups are bonded, these are non-reducing sugars e.g. sucrose. On the other hand,

sugars in which these functional groups are free, are called reducing sugars, for example, maltose and lactose.

9.1.2 Monosaccharides

Monosaccharides are further classified on the basis of number of carbon atoms and the functional group present in them. If a monosaccharide contains an aldehyde group, it is known as an aldose and if it contains a keto group, it is known as a ketose. Number of carbon atoms constituting the monosaccharide is also introduced in the name as is evident from the examples given in Table 9.1

Table 9.1: Different Types of Monosaccharides

Carbon atoms	General term	Aldehyde	Ketone
3	Triose	Aldotriose	Ketotriose
4	Tetrose	Aldotetrose	Ketotetrose
5	Pentose	Aldopentose	Ketopentose
6	Hexose	Aldohexose	Ketohexose
7	Heptose	Aldoheptose	Ketoheptose

I Glucose

Glucose occurs freely in nature as well as in the combined form. It is present in sweet fruits and honey. Ripe grapes also contain glucose in large amounts. It is prepared as follows:

9.1.3 Preparation of Glucose

1. *From sucrose (Cane sugar)*: If sucrose is boiled with dilute HCl or H_2SO_4 in alcoholic solution, glucose and fructose are obtained in equal amounts.

$$\underset{\text{Sucrose}}{C_{12}H_{22}O_{11}} + H_2O \xrightarrow{H^+} \underset{\text{Glucose}}{C_6H_{12}O_6} + \underset{\text{Fructose}}{C_6H_{12}O_6}$$

2. *From starch*: Commercially glucose is obtained by hydrolysis of starch by boiling it with dilute H_2SO_4 at 393 K under pressure.

$$\underset{\text{Starch or cellulose}}{(C_6H_{10}O_5)_n} + nH_2O \xrightarrow[393\text{K; 2-3 atm}]{H^+} \underset{\text{Glucose}}{nC_6H_{12}O_6}$$

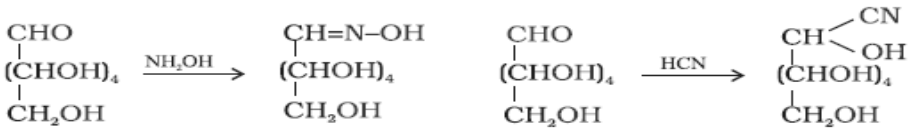
9.1.4 Structure of Glucose

Glucose is an aldohexose and is also known as dextrose. It is the monomer of many of the larger carbohydrates, namely starch, cellulose. It is probably the most abundant organic compound on earth. It was assigned the structure given below on the basis of the following evidences:

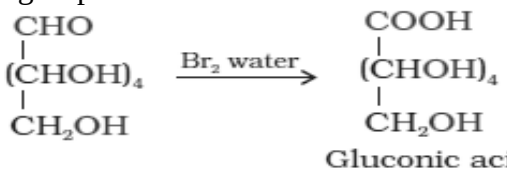
1. Its molecular formula was found to be $C_6H_{12}O_6$.
2. On prolonged heating with HI, it forms n-hexane, suggesting that all the six carbon atoms are linked in a straight chain.

$$\begin{array}{l} CHO \\ | \\ (CHOH)_4 \\ | \\ CH_2OH \end{array} \xrightarrow{HI,\ \Delta} \underset{(n\text{-Hexane})}{CH_3{-}CH_2{-}CH_2{-}CH_2{-}CH_2{-}CH_3}$$

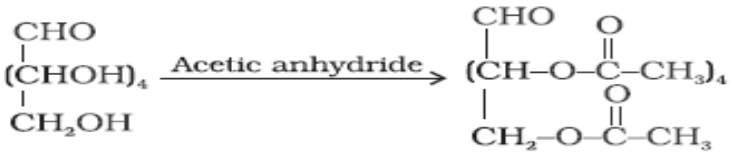
3. Glucose reacts with hydroxylamine to form an oxime and adds a molecule of hydrogen cyanide to give cyanohydrin. These reactions confirm the presence of a carbonyl group (>C = 0) in glucose.



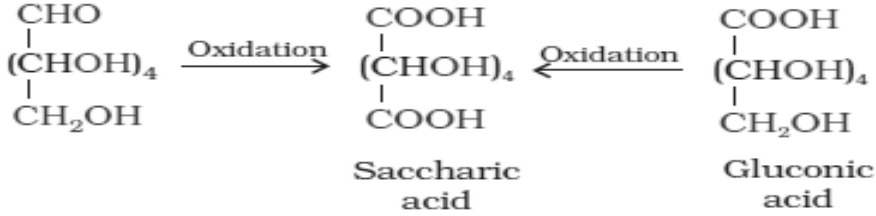
4. Glucose gets oxidised to six carbon carboxylic acid (gluconic acid) on reaction with a mild oxidising agent like bromine water. This indicates that the carbonyl group is present as an aldehydic group.



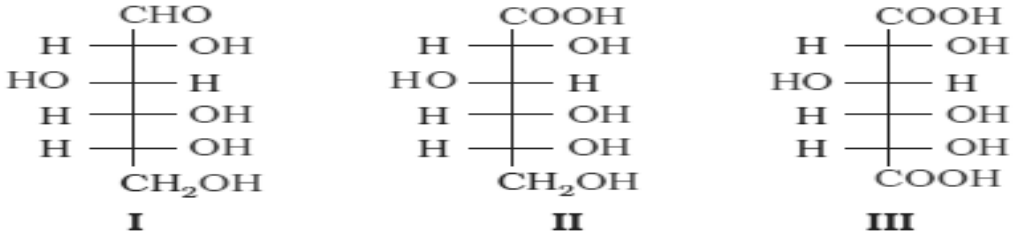
5. Acetylation of glucose with acetic anhydride gives glucose pentaacetate which confirms the presence of five –OH groups. Since it exists as a stable compound, five –OH groups should be attached to different carbon atoms.



6. On oxidation with nitric acid, glucose as well as gluconic acid both yield a dicarboxylic acid, saccharic acid. This indicates the presence of a primary alcoholic (–OH) group in glucose.

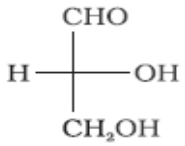


The exact spatial arrangement of different —OH groups was given by Fischer after studying many other properties. Its configuration is correctly represented as **I**. So gluconic acid is represented as **II** and saccharic acid as **III**.

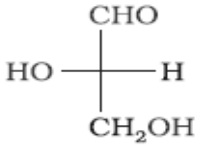


Glucose is correctly named as D(+)-glucose. 'D' before the name of glucose represents the configuration whereas '(+)' represents dextrorotatory nature of the molecule. It may be remembered that 'D' and 'L' have no relation with the optical activity of the compound.

The meaning of D– and L– notations is given as follows. The letters 'D' or 'L' before the name of any compound indicate the relative configuration of a particular stereoisomer. This refers to their relation with a particular isomer of glyceraldehyde. Glyceraldehyde contains one asymmetric carbon atom and exists in two enantiomeric forms as shown below.

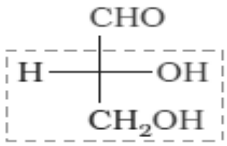


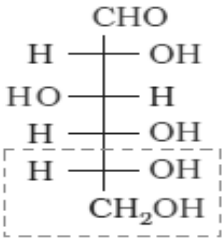
(+) – Glyceraldehyde



(–) – Glyceraldehyde

All those compounds which can be chemically correlated to (+) isomer of glyceraldehyde are said to have D-configuration whereas those which can be correlated to (–) isomer of glyceraldehyde are said to have L—configuration. For assigning the configuration of monosaccharides, it is the lowest asymmetric carbon atom (as shown below) which is compared. As in (+) glucose, —OH on the lowest asymmetric carbon is on the right side which is comparable to (+) glyceraldehyde, so it is assigned D-configuration. For this comparison, the structure is written in a way that most oxidised carbon is at the top.





9.1.5 Cyclic Structure of Glucose

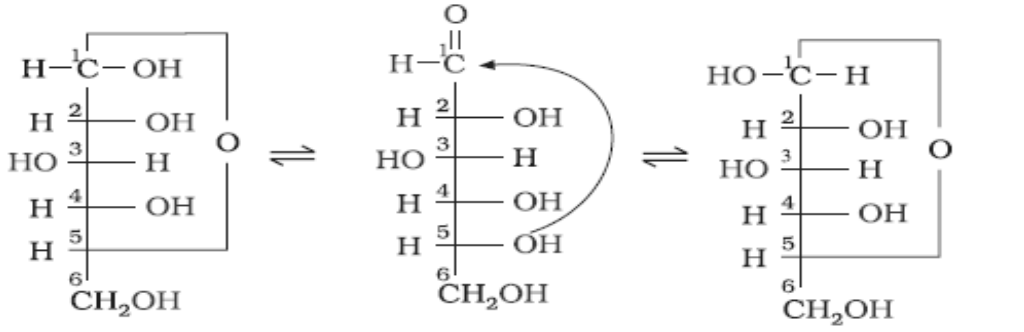
The structure **(I)** of glucose explained most of its properties but the following reactions and facts could not be explained by this structure.

1. Despite having the aldehyde group, glucose does not give 2,4-DNP test, Schiff's test and it does not form the hydrogensulphite addition product with $NaHSO_3$.
2. The pentaacetate of glucose does not react with hydroxylamine indicating the absence of free —CHO group.
3. Glucose is found to exist in two different crystalline forms which are named as α and β. The α-form of glucose (m.p. 419 K) is obtained by crystallisation from concentrated solution of glucose at 303 K while the β-form (m.p. 423 K) is obtained by crystallisation from hot and saturated aqueous solution at 371 K.

This behaviour could not be explained by the open chain structure

(**I**) for glucose. It was proposed that one of the —OH groups may add to the —CHO group and form a cyclic hemiacetal structure. It was found that glucose forms a six-membered ring in which —OH at C-5 is involved in ring formation. This explains the absence of —CHO group and also existence of glucose in two forms as shown below.

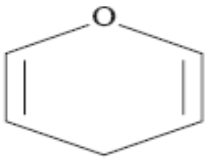
These two cyclic forms exist in equilibrium with open chain structure.



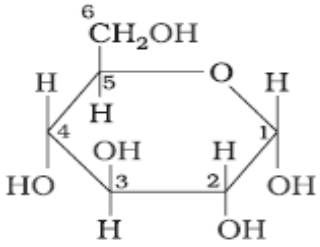
α – D – (+) – Glucose

β – D– (+) – Glucose

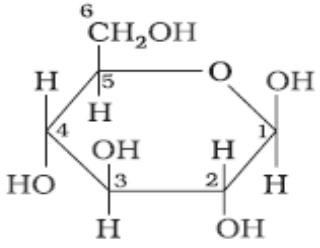
The two cyclic hemiacetal forms of glucose differ only in the configuration of the hydroxyl group at C1, called *anomeric carbon* (the aldehyde carbon before cyclisation). Such isomers, i.e., α- form and β-form, are called **anomers**. The six membered cyclic structure of glucose is called **pyranose structure** (α- or β-form), in analogy with pyran. Pyran is a cyclic organic compound with one oxygen atom and five carbon atoms in the ring. The cyclic structure of glucose is more correctly represented by Haworth structure as given below.



Pyran



α – D – (+) – Glucopyranose



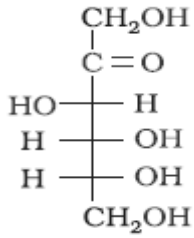
β – D – (+) – Glucopyranose

II. Fructose

Fructose is an important ketohexose. It is obtained along with glucose by the hydrolysis of disaccharide, sucrose.

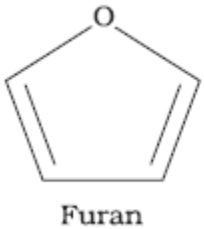
9.1.6 Structure of Fructose

Fructose also has the molecular formula $C_6H_{12}O_6$ and on the basis of its reactions it was found to contain a ketonic functional group at carbon number 2 and six carbons in straight chain as in the case of glucose. It belongs to D-series and is a laevorotatory compound. It is appropriately written as D-(–)-fructose. Its open chain structure is as shown.

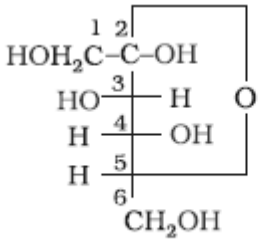


D – (–) – Fructose

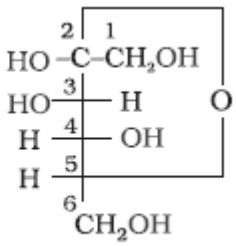
It also exists in two cyclic forms which are obtained by the addition of —OH at C_5 to the ($>C=O$) group. The ring, thus formed is a five membered ring and is named as furanose with analogy to the compound furan. Furan is a five membered cyclic compound with one oxygen and four carbon atoms.



Furan

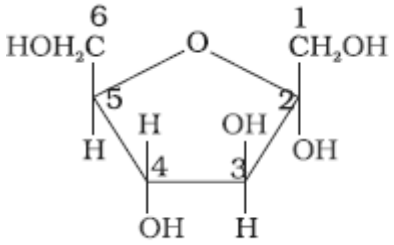


α – D – (–) – Fructofuranose

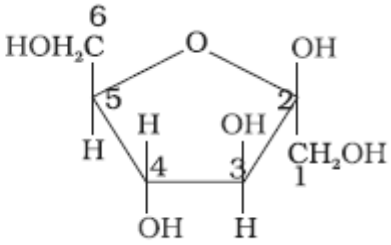


β – D – (–) – Fructofuranose

The cyclic structures of two anomers of fructose are represented by Haworth structures as given.



α – D – (–) – Fructofuranose



β – D – (–) – Fructofuranose

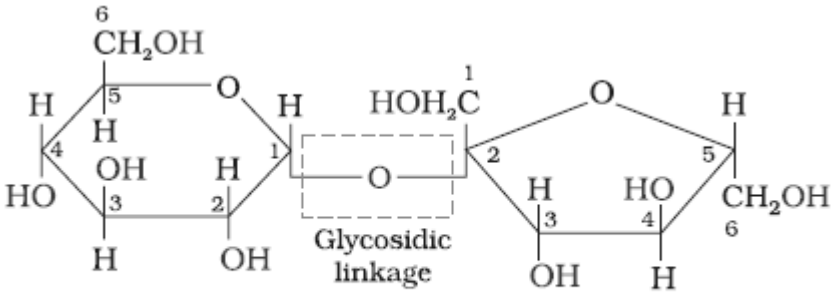
9.1.7 Disaccharides

You have already read that disaccharides on hydrolysis with dilute acids or enzymes yield two molecules of either the same or different monosaccharides. The two monosaccharides are joined together by an oxide linkage formed by the loss of a water molecule. Such a linkage between two monosaccharide units through oxygen atom is called *glycosidic linkage.*

(i) *Sucrose***:** One of the common disaccharides is **sucrose** which on hydrolysis gives equimolar mixture of D-(+)-glucose and D-(-) fructose.

$$\underset{\text{Sucrose}}{C_{12}H_{22}O_{11}} + H_2O \longrightarrow \underset{\text{D-(+)-Glucose}}{C_6H_{12}O_6} + \underset{\text{D-(–)-Fructose}}{C_6H_{12}O_6}$$

These two monosaccharides are held together by a glycosidic linkage between C_1 of α-glucose and C_2 of β-fructose. Since the reducing groups of glucose and fructose are involved in glycosidic bond formation, sucrose is a non reducing sugar.

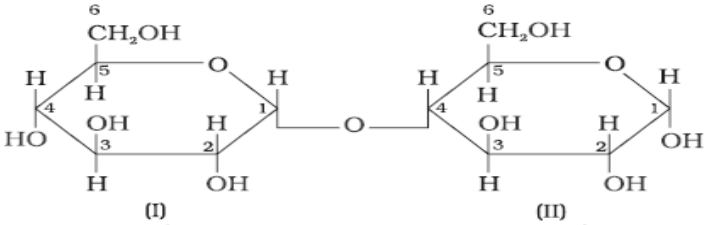


α – D – Glucose β – D – Fructose

Sucrose

Sucrose is dextrorotatory but after hydrolysis gives dextrorotatory glucose and laevorotatory fructose. Since the laevorotation of fructose (–92.4°) is more than dextrorotation of glucose (+ 52.5°), the mixture is laevorotatory. Thus, hydrolysis of sucrose brings about a change in the sign of rotation, from dextro (+) to laevo (–) and the product is named as **invert sugar**.

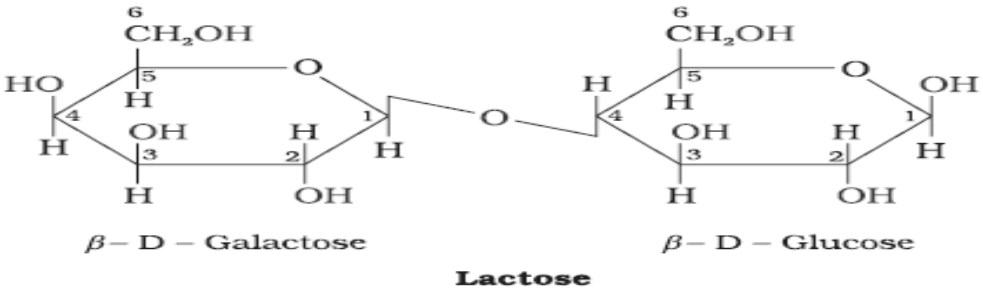
(ii) *Maltose*: Another disaccharide, maltose is composed of two α-D-glucose units in which C_1 of one glucose (I) is linked to C_4 of another glucose unit (II). The free aldehyde group can be produced at C_1 of second glucose in solution and it shows reducing properties so it is a reducing sugar.



(I) α – D – Glucose (II) α – D – Glucose

Maltose

(iii) *Lactose*: It is more commonly known as milk sugar since this disaccharide is found in milk. It is composed of β-D-galactose and β-D-glucose. The linkage is between C1 of galactose and C4 of glucose. Hence it is also a reducing sugar.



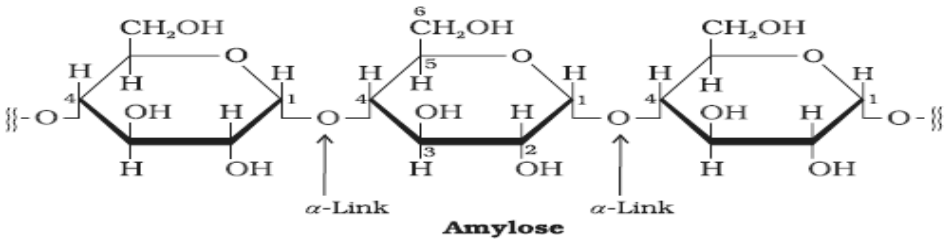
Lactose

9.1.8 Polysaccharides

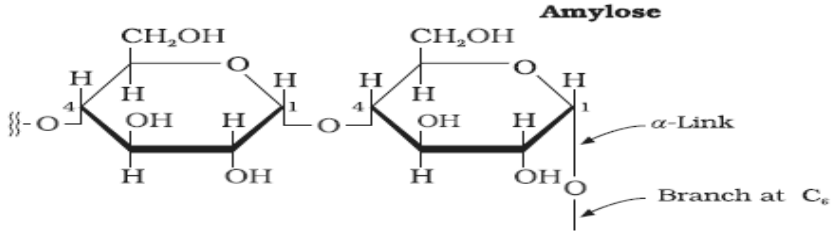
Polysaccharides contain a large number of monosaccharide units joined together by glycosidic linkages. These are the most commonly encountered carbohydrates in nature. They mainly act as the food storage or structural materials.

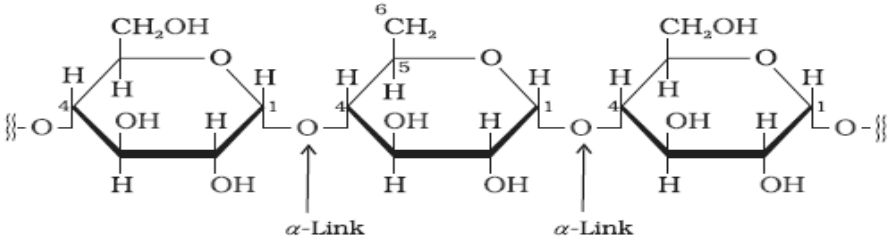
(i) *Starch*: Starch is the main storage polysaccharide of plants. It isthe most important dietary source for human beings. High content of starch is found in cereals, roots, tubers and some vegetables. It is a polymer of α-glucose and consists of two components— **Amylose** and **Amylopectin**. Amylose is water soluble component which constitutes about 15-20% of starch. Chemically amylose is a long unbranched chain with 200-1000 α-D-(+)-glucose units held by C_1– C_4 glycosidic linkage.

Amylopectin is insoluble in water and constitutes about 80-85% of starch. It is a branched chain polymer of α-D-glucose units in which chain is formed by C_1–C_4 glycosidic linkage whereas branching occurs by C_1–C_6 glycosidic linkage.



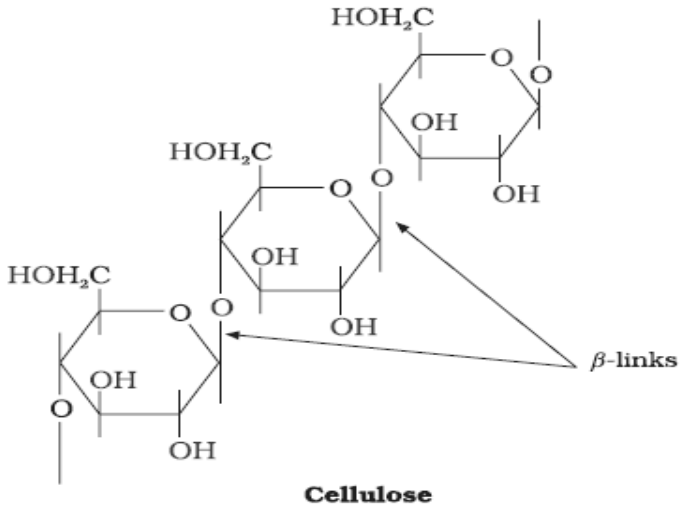
Amylose





Amylopectin

(ii) *Cellulose*: Cellulose occurs exclusively in plants and it is the most abundant organic substance in plant kingdom. It is a predominant constituent of cell wall of plant cells. Cellulose is a straight chain



Cellulose

polysaccharide composed only of β-D-glucose units which are joined by glycosidic linkage between C_1 of one glucose unit and C_4 of the next glucose unit.

(iii) *Glycogen*: The carbohydrates are stored in animal body as glycogen. It is also known as *animal starch* because its structure is similar to amylopectin and is rather more highly branched. It is present in liver, muscles and brain. When the body needs glucose, enzymes break the glycogen down to glucose. Glycogen is also found in yeast and fungi.

9.1.9 Importance of Carbohydrates

Carbohydrates are essential for life in both plants and animals. They form a major portion of our food. Honey has been used for a long time as an instant source of energy by '**Vaids**' in ayurvedic system of medicine. Carbohydrates are used as storage molecules as starch in plants and **glycogen** in animals. Cell wall of bacteria and plants is made up of cellulose. We build furniture, etc. from cellulose in the form of wood and clothe ourselves with cellulose in the form of cotton fibre. They provide raw materials for many important industries like textiles, paper, lacquers and breweries.

Two aldopentoses viz. D-ribose and 2-deoxy-D-ribose (Section 9.5.1, Class XII) are present in nucleic acids. Carbohydrates are found in biosystem in combination with many proteins and lipids.

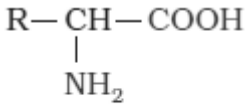
9.2 Proteins

Proteins are the most abundant biomolecules of the living system.Chief sources of proteins are milk, cheese, pulses, peanuts, fish, meat, etc. They occur in every part of the body and form the fundamental basis of structure and functions of life. They are also required for growth and maintenance of body. The word protein is derived from Greek word, "**proteios**" which means primary or of prime importance. All proteins are polymers of α-amino acids.

9.2.1 Amino Acids

Amino acids contain amino ($–NH_2$) and carboxyl (–COOH) functional groups. Depending upon the relative position of amino group with respect to carboxyl group, the

amino acids can be classified as α,β, γ, δ and so on. Only α-amino acids are obtained on hydrolysis of proteins. They may contain other functional groups also.



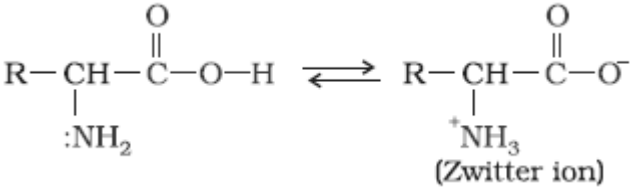
α-amino acid
(R = side chain)

All α-amino acids have trivial names, which usually reflect the property of that compound or its source. Glycine is so named since it has sweet taste (in Greek *glykos* means sweet) and tyrosine was first obtained from cheese (in Greek, *tyros* means cheese.) Amino acids are generally represented by a three letter symbol, sometimes one letter symbol is also used. Structures of some commonly occurring amino acids along with their 3-letter and 1-letter symbols are given in Table 9.2.

9.2.2 Classification of Amino Acids

Amino acids are classified as acidic, basic or neutral depending upon the relative number of amino and carboxyl groups in their molecule. Equal number of amino and carboxyl groups makes it neutral; more number of amino than carboxyl groups makes it basic and more carboxyl groups as compared to amino groups makes it acidic. The amino acids, which can be synthesised in the body, are known as **nonessential amino acids.** On the other hand, those which cannot be synthesised in the body and must be obtained through diet, are known as **essential amino acids** (marked with asterisk in Table 9.2).

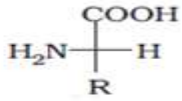
Amino acids are usually colourless, crystalline solids. These are water-soluble, high melting solids and behave like salts rather than simple amines or carboxylic acids. This behaviour is due to the presence of both acidic (carboxyl group) and basic (amino group) groups in the same molecule. In aqueous solution, the carboxyl group can lose a proton and amino group can accept a proton, giving rise to a dipolar ion known as *zwitter ion*. This is neutral but contains both positive and negative charges.

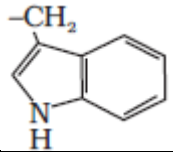
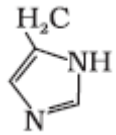
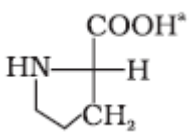


(Zwitter ion)

In zwitter ionic form, amino acids show amphoteric behaviour as they react both with acids and bases.

Except glycine, all other naturally occurring α-amino acids are optically active, since the α-carbon atom is asymmetric. These exist both in 'D' and 'L' forms. Most naturally occurring amino acids have L- configuration. L-Aminoacids are represented by writing the $–NH_2$ group on left hand side.

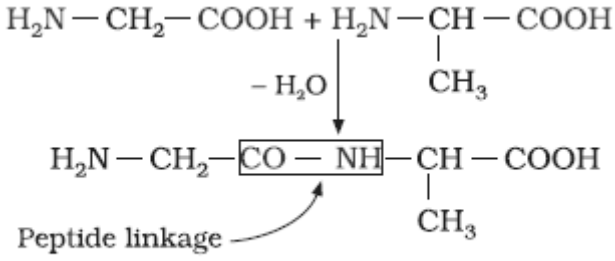
Table 9.2 : Natural Amino Acids 

Name of the amino acids	Characteristic feature of side chain, R	Three letter symbol	One letter code
1. Glycine	H	Gly	G
2. Alanine	$-CH_3$	Ala	A
3. Valine*	$(H_3C)_2CH-$	Val	V
4. Leucine*	$(H_3C)_2CH-CH_2-$	Leu	L
5. Isoleucine*	$H_3C-CH_2-CH(CH_3)-$	Ile	I
6. Arginine*	$HN=C(NH_2)-NH-(CH_2)_3-$	Arg	R
7. Lysine*	$H_2N-(CH_2)_4-$	Lys	K
8. Glutamic acid	$HOOC-CH_2-CH_2-$	Glu	E
9. Aspartic acid	$HOOC-CH_2-$	Asp	D
10. Glutamine	$H_2N-C(=O)-CH_2-CH_2-$	Gln	Q
11. Asparagine	$H_2N-C(=O)-CH_2-$	Asn	N
12. Threonine*	$H_3C-CHOH-$	Thr	T
13. Serine	$HO-CH_2-$	Ser	S
9. Cysteine	$HS-CH_2-$	Cys	C
15. Methionine*	$H_3C-S-CH_2-CH_2-$	Met	M
16. Phenylalanine*	$C_6H_5-CH_2-$	Phe	F
17. Tyrosine	$(p)HO-C_6H_4-CH_2-$	Tyr	Y
18. Tryptophan*		Trp	W
19. Histidine*		His	H
20. Proline		Pro	P

** essential amino acid, a = entire structure*

9.2.3 Structure of Proteins

You have already read that proteins are the polymers of α-amino acids and they are connected to each other by **peptide bond** or **peptide linkage.** Chemically, peptide linkage is an amide formed between –COOH group and $–NH_2$ group. The reaction between two molecules of similar or different amino acids, proceeds through the combination of the amino group of one molecule with the carboxyl group of the other. This results in the elimination of a water molecule and formation of a peptide bond –CO–NH–. The product of the reaction is called a dipeptide because it is made up of two amino acids. For example, when carboxyl group of glycine combines with the amino group of alanine we get a **dipeptide**, glycylalanine.



Glycylalanine (Gly-Ala)

If a third amino acid combines to a dipeptide, the product is called a **tripeptide**. A tripeptide contains three amino acids linked by two peptide linkages. Similarly when four, five or six amino acids are linked, the respective products are known as **tetrapeptide, pentapeptide or hexapeptide**, respectively. When the number of such amino acids is more than ten, then the products are called **polypeptides**. A polypeptide with more than hundred amino acid residues, having molecular mass higher than 10,000u is called a protein. However, the distinction between a polypeptide and a protein is not very sharp. Polypeptides with fewer amino acids are likely to be called proteins if they ordinarily have a well defined conformation of a protein such as insulin which contains 51 amino acids.

Proteins can be classified into two types on the basis of their molecular shape.

(a) Fibrous proteins

When the polypeptide chains run parallel and are held together by hydrogen and disulphide bonds, then fibre– like structure is formed. Such proteins are generally insoluble in water. Some common examples are keratin (present in hair, wool, silk) and myosin (present in muscles), etc.

(b) Globular proteins

This structure results when the chains of polypeptides coil around to give a spherical shape. These are usually soluble in water. Insulin and albumins are the common examples of globular proteins.

Structure and shape of proteins can be studied at four different levels, i.e., primary, secondary, tertiary and quaternary, each level being more complex than the previous one.

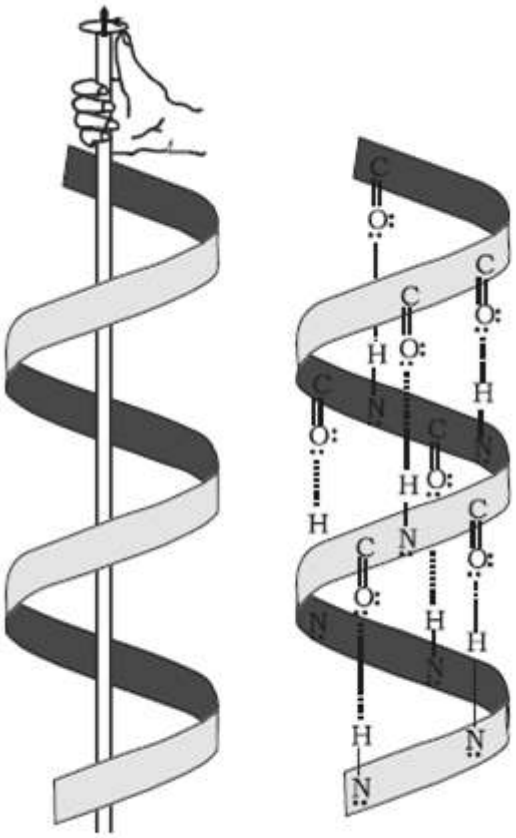


Fig 9.1 α–Helix structure of proteins

(i) *Primary structure of proteins*: Proteins may have one or more polypeptide chains. Each polypeptide in a protein has amino acids linked with each other in a specific sequence and it is this sequence of amino acids that is said to be the primary structure of that protein. Any change in this primary structure i.e., the sequence of amino acids creates a different protein.

(ii) *Secondary structure of proteins***:** The secondary structure of protein refers to the shape in which a long polypeptide chain can exist. They are found to exist in two different types of structures viz. α-helix and β-pleated sheet structure. These structures arise due to the regular folding of the backbone of the polypeptide chain due to hydrogen bonding between $-\overset{\overset{O}{\|}}{C}-$ and –NH– groups of the peptide bond. □-Helix is one of the most common ways in which a polypeptide chain forms all possible hydrogen bonds by twisting into a right handed screw (helix) with the –NH group of each amino acid residue hydrogen bonded to the $>\overset{-}{C}=O$ of an adjacent turn of the helix as shown in Fig. 9.1.

In β-structure all peptide chains are stretched out to nearly maximum extension and then laid side by side which are held together by intermolecular hydrogen bonds. The structure resembles the pleated folds of drapery and therefore is known as β-pleated sheet.

(iii) *Tertiary structure of proteins***:** The tertiary structure of proteins represents overall folding of the polypeptide chains i.e., further folding of the secondary structure. It gives rise to two major molecular shapes viz. fibrous and globular. The main forces which stabilise the 2° and 3° structures of proteins are hydrogen bonds, disulphide linkages, Vander Waals and electrostatic forces of attraction.

(iv) Quaternary structure of proteins: Some of the proteins are composed of two or more polypeptide chains referred to as sub-units. The spatial arrangement of these subunits with respect to each other is known as quaternary structure.

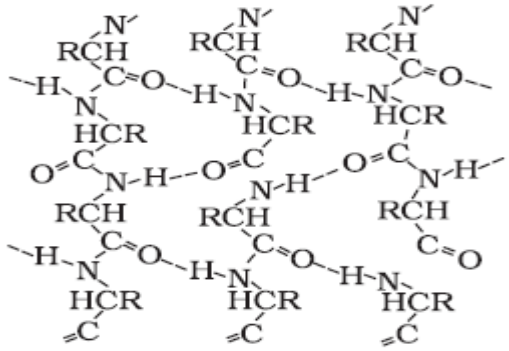


Fig 9.2 β–Pleated sheet structure of proteins

A diagrammatic representation of all these four structures is given in Figure 9.3 where each coloured ball represents an amino acid.

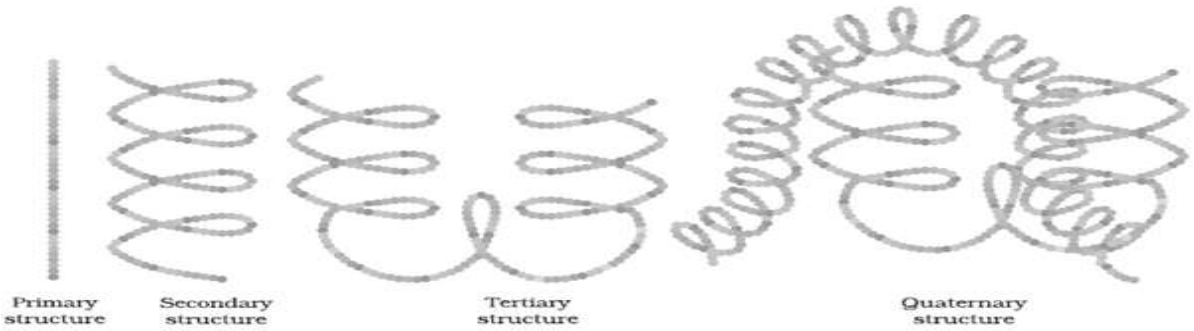


Fig 9.3 Diagramatic representation of protein structure (two sub units of two types in quaternanry structure)

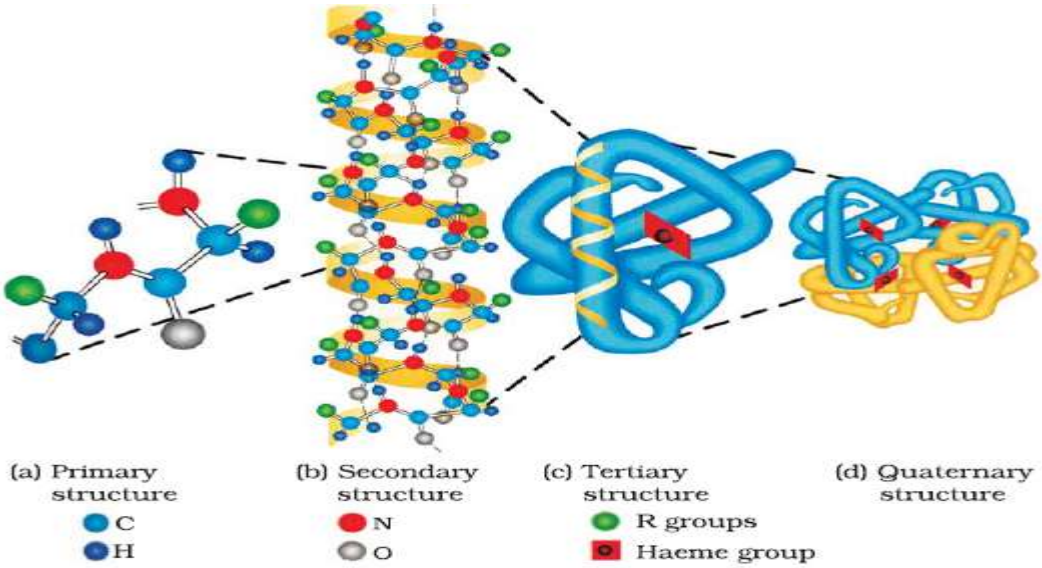


Fig. 9.4 Primary,secondary, tertiaryand quaternarystructures ofhaemoglobin

9.2.4 Denaturation of Proteins

Protein found in a biological system with a unique three-dimensional structure and biological activity is called a native protein. When a protein in its native form, is subjected to physical change like change in temperature or chemical change like change in pH, the hydrogen bonds are disturbed. Due to this, globules unfold and helix get uncoiled and protein loses its biological activity. This is called **denaturation** of protein. During denaturation 2° and 3° structures are destroyed but 1º structure remains intact. The coagulation of egg white on boiling is a common example of denaturation. Another example is curdling of milk which is caused due to the formation of lactic acid by the bacteria present in milk.

9.3 Enzymes

Life is possible due to the coordination of various chemical reactions in living organisms. An example is the digestion of food, absorption of appropriate molecules and ultimately production of energy. This process involves a sequence of reactions and all these reactions occur in the body under very mild conditions. This occurs with the help of certain biocatalysts called **enzymes.** Almost all the enzymes are globular proteins. Enzymes are very specific for a particular reaction and for a particular substrate. They are generally named after the compound or class of compounds upon which they work. For example, the enzyme that catalyses hydrolysis of maltose into glucose is named as *maltase*.

$$\underset{\text{Maltose}}{C_{12}H_{22}O_{11}} \xrightarrow{\text{Maltase}} 2\,\underset{\text{Glucose}}{C_6H_{12}O_6}$$

Sometimes enzymes are also named after the reaction, where they are used. For example, the enzymes which catalyse the oxidation of one substrate with simultaneous reduction of another substrate are named as **oxidoreductase** enzymes. The ending of the name of an enzyme is **-ase**.

9.3.1 Mechanism of Enzyme Action

Enzymes are needed only in small quantities for the progress of a reaction.Similar to the action of chemical catalysts, enzymes are said to reduce the magnitude of activation energy. For example, activation energy for acid hydrolysis of sucrose is 6.22 kJ mol^{-1}, while the activation energy is only 2.15 kJ mol^{-1} when hydrolysed by the enzyme, sucrase. Mechanism for the enzyme action has been discussed in Unit 5.

9.4 Vitamins

It has been observed that certain organic compounds are required in small amounts in our diet but their deficiency causes specific diseases. These compounds are called **vitamins.** Most of the vitamins cannot be synthesised in our body but plants can synthesise almost all of them, so they are considered as essential food factors. However, the bacteria of the gut can produce some of the vitamins required by us. All the vitamins are generally available in our diet. Different vitamins belong to various chemical classes and it is difficult to define them on the basis of structure. They are generally regarded as **organic compounds required in the diet in small amounts to perform specific biological functions for normal aintenance of optimum growth and health of the organism.** Vitamins are designated by alphabets A, B, C, D, etc. Some of them are further named as sub-groups e.g. B1, B2, B6, B12, etc. Excess of vitamins is also harmful and vitamin pills should not be taken without the advice of doctor.

The term "**Vitamine**" was coined from the word vital + amine since the earlier identified compounds had amino groups. Later work showed that most of them did not contain amino groups, so the letter 'e' was dropped and the term **vitamin** is used these days.

9.4.1 Classification of Vitamins

Vitamins are classified into two groups depending upon their solubility in water or fat.

(i) *Fat soluble vitamins*: Vitamins which are soluble in fat and oils but insoluble in water are kept in this group. These are vitamins A, D, E and K. They are stored in liver and adipose (fat storing) tissues.

(ii) *Water soluble vitamins*: B group vitamins and vitamin C are soluble in water so they are grouped together. Water soluble vitamins must be supplied regularly in diet because they are readily excreted in urine and cannot be stored (except vitamin B12) in our body.

Some important vitamins, their sources and diseases caused by their deficiency are listed in Table 9.3.

Table 9.3 Some important vitamins, their sources and diseases caused by their deficiency

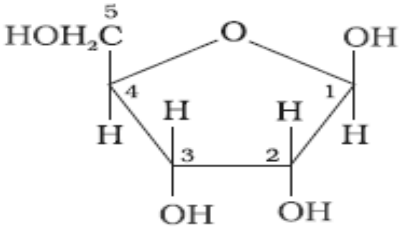
Sl. No.	Name of Vitamins	Sources	Deficiency diseases
1.	Vitamin A	Fish liver oil, carrots, butter and milk	Xerophthalmia (hardening of cornea of eye) Night blindness
2.	Vitamin B_1 (Thiamine)	Yeast, milk, green vegetables and cereals	Beri beri (loss of appetite, retarded growth)
3.	Vitamin B_2 (Riboflavin)	Milk, eggwhite, liver, kidney	Cheilosis (fissuring at corners of mouth and lips), digestive disorders and burning sensation of the skin.
4.	Vitamin B_6 (Pyridoxine)	Yeast, milk, egg yolk, cereals and grams	Convulsions
5.	Vitamin B_{12}	Meat, fish, egg and curd	Pernicious anaemia (RBC deficient in haemoglobin)
6.	Vitamin C (Ascorbic acid)	Citrus fruits, amla and green leafy vegetables	Scurvy (bleeding gums)
7.	Vitamin D	Exposure to sunlight, fish and egg yolk	Rickets (bone deformities in children) and osteomalacia (soft bones and joint pain in adults)
8.	Vitamin E	Vegetable oils like wheat germ oil, sunflower oil, etc.	Increased fragility of RBCs and muscular weakness
9.	Vitamin K	Green leafy vegetables	Increased blood clotting time

9.5: Nucleic Acids

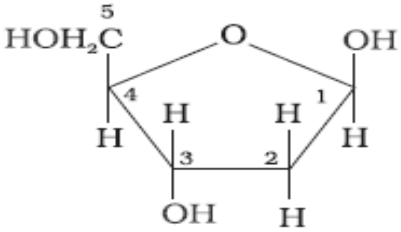
Every generation of each and every species resembles its ancestors in many ways. How are these characteristics transmitted from one generation to the next? It has been observed that nucleus of a living cell is responsible for this transmission of inherent characters, also called **heredity**. The particles in nucleus of the cell, responsible for heredity, are called chromosomes which are made up of proteins and another type of bio-molecules called **nucleic acids.** These are mainly of two types, the **deoxyribonucleic acid (DNA) and ribonucleic acid (RNA).** Since nucleic acids are long chain polymers of **nucleotides,** so they are also called poly-nucleotides.

9.5.1 Chemical Composition of Nucleic Acids

Complete hydrolysis of DNA (or RNA) yields a pentose sugar, phosphoric acid and nitrogen containing heterocyclic compounds (called bases). In DNA molecules, the sugar moiety is β-D-2-deoxyribose whereas in RNA molecule, it is β-D-ribose.

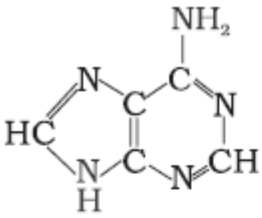


β–D–ribose

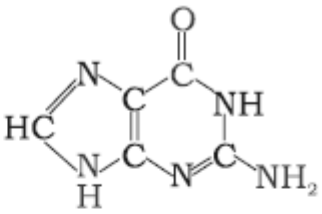


β–D–2–deoxyribose

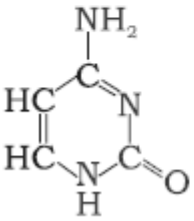
DNA contains four bases viz. adenine (A), guanine (G), cytosine (C) and thymine (T). RNA also contains four bases, the first three bases aresame as in DNA but the fourth one is uracil (U).



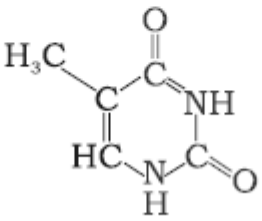
Adenine (A)



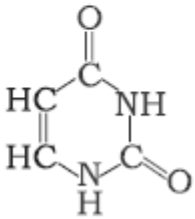
Guanine (G)



Cytosine (C)



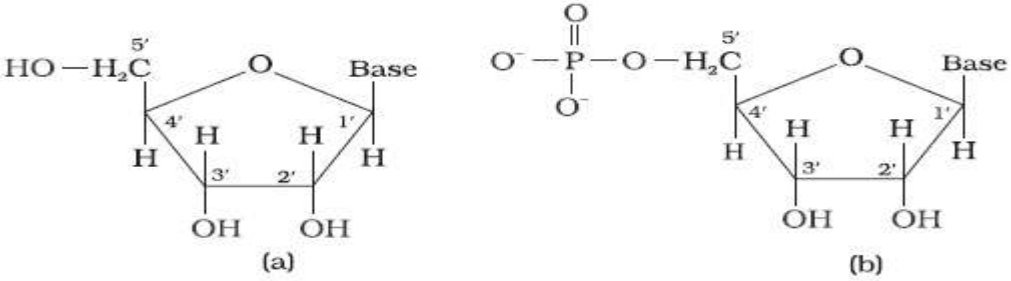
Thymine (T)



Uracil (U)

9.5.2 Structure of Nucleic Acids

A unit formed by the attachment of a base to 1′ position of sugar is known as **nucleoside.** In nucleosides, the sugar carbons are numbered as 1′, 2′, 3′, etc. in order to distinguish these from the bases (Fig. 9.5a). When nucleoside is linked to phosphoric acid at 5′-position of sugar moiety, we get a nucleotide (Fig. 9.5b).



(a) (b)

Fig. **9.5** : *Structure of (a) a nucleoside and (b) a nucleotide*

Nucleotides are joined together by phosphodiester linkage between 5′ and 3′ carbon atoms of the pentose sugar. The formation of a typical dinucleotide is shown in Fig. 9.6.

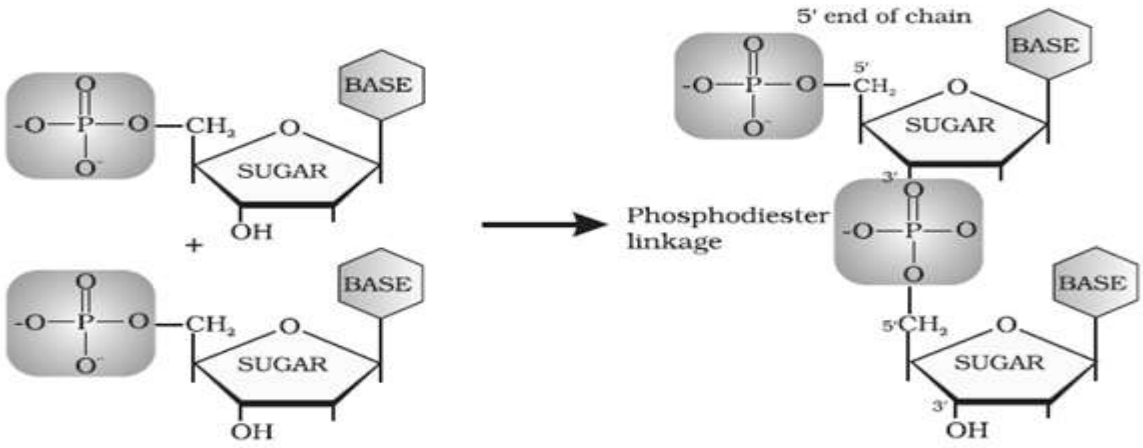
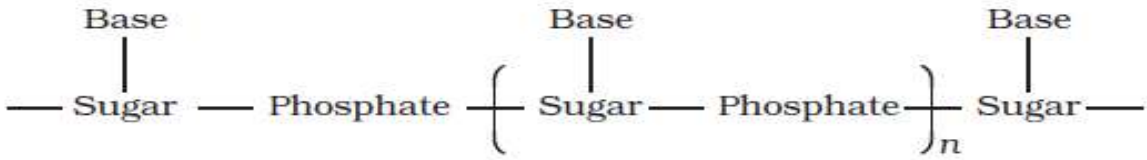


Fig. 9.6 Formation of a dinucleotide

A simplified version of nucleic acid chain is as shown below.



Information regarding the sequence of nucleotides in the chain of a nucleic acid is called its primary structure. Nucleic acids have a secondary structure also. James Watson and Francis Crick gave a double strand helix structure for DNA (Fig. 9.7). Two nucleic acid chains are wound about each other and held together by hydrogen bonds between pairs of bases. The two strands are complementary to each other because the hydrogen bonds are formed between specific pairs of bases. Adenine forms hydrogen bonds with thymine whereas cytosine forms hydrogen bonds with guanine.

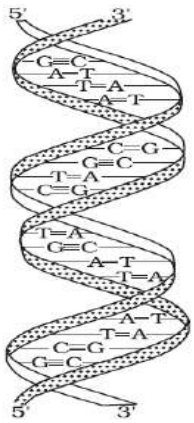


Fig. 9.7: Double strand helix structure for DNA

In secondary structure of RNA, helices are present which are only single stranded. Sometimes they fold back on themselves to form a double helix structure. RNA molecules are of three types and they perform different functions. They are named as **messenger RNA (m-RNA), ribosomal RNA (r-RNA) and transfer RNA (t-RNA).**

9.5.3 Biological Functions of Nucleic Acids

DNA is the chemical basis of heredity and may be regarded as the reserve of genetic information. DNA is exclusively responsible for maintaining the identity of different species of organisms over millions of years. A DNA molecule is capable of self duplication during cell division and identical DNA strands are transferred to daughter cells. Another important function of nucleic acids is the protein synthesis in the cell. Actually, the proteins are synthesised by various RNA molecules in the cell but the message for the synthesis of a particular protein is present in DNA.

Summary

Carbohydrates are optically active polyhydroxy aldehydes or ketones or molecules which provide such units on hydrolysis. They are broadly classified into three groups — **monosaccharides**, **disaccharides** and **polysaccharides**. Glucose, the most important source of energy for mammals, is obtained by the digestion of starch. Monosaccharides are held together by glycosidic linkages to form disaccharides or polysaccharides.

Proteins are the **polymers** of about twenty different **α-amino acids** which are linked by peptide bonds. Ten amino acids are called essential amino acids because they cannot be synthesised by our body, hence must be provided through diet. Proteins perform various structural and dynamic functions in the organisms. Proteins which contain only α-amino acids are called simple proteins. The **secondary** or **tertiary structure of proteins** get disturbed on change of pH or temperature and they are not able to perform their functions. This is called **denaturation of proteins**. Enzymes are **biocatalysts** which speed up the reactions in biosystems. They are very specific and selective in their action and chemically all **enzymes** are proteins.

Vitamins are accessory food factors required in the diet. They are classified as fat soluble (A, D, E and K) and water soluble (B group and C). Deficiency of vitamins leads to many diseases.

Nucleic acids are the polymers of nucleotides which in turn consist of a base, a pentose sugar and phosphate moiety. Nucleic acids are responsible for the transfer of characters from parents to offsprings. There are two types of nucleic acids — **DNA** and **RNA**. DNA contains a five carbon sugar molecule called **2-deoxyribose** whereas RNA contains ribose. Both DNA and RNA contain adenine, guanine and cytosine. The fourth base is thymine in DNA and uracil in RNA. The structure of DNA is a double strand whereas RNA is a single strand molecule. DNA is the chemical basis of heredity and have the coded message for proteins to be synthesised in the cell. There are three types of RNA — mRNA, rRNA and tRNA which actually carry out the protein synthesis in the cell.

Important Questions

1. Give the sources of the following vitamins and name the diseases caused by their deficiency
    a. Water Soluble Vitamins
    b. Fat Soluble Vitamins
2. What are hormones? Give one example for each.
    a. Steroid Hormones
    b. Polypeptide Hormones
    c. Amino-acid derivatives
3. Differentiate between globular and fibrous proteins
4. What are essential and non-essential Amino-acids? Give one example of each.
5. What is Zwitter ion? Give an example.

CHAPTER 10

Chemistry in Everyday Life

The principles of chemistry have been used for the benefit of mankind. Think of cleanliness — the materials like soaps, detergents, household bleaches, tooth pastes, etc. will come to your mind. Look towards the beautiful clothes — immediately chemicals of the synthetic fibres used for making clothes and chemicals giving colours to them will come to your mind. Food materials — again a number of chemicals about which you have learnt in the previous Unit will appear in your mind. Of course, sickness and diseases remind us of medicines — again chemicals. Explosives, fuels, rocket propellants, building and electronic materials, etc., are all chemicals. Chemistry has influenced our life so much that we do not even realise that we come across chemicals at every moment; that we ourselves are beautiful chemical creations and all our activities are controlled by chemicals. In this Unit, we shall learn the application of Chemistry in three important and interesting areas, namely – medicines, food materials and cleansing agents.

10.1 Drugs and their Classification

Drugs are chemicals of low molecular masses (~100–500u). These interact with macromolecular targets and produce a biological response. When the biological response is therapeutic and useful, these chemicals are called **medicines** and are used in diagnosis, prevention and treatment of diseases. If taken in doses higher than those recommended, most of the drugs used as medicines are potential poisons. Use of chemicals for therapeutic effect is called **chemotherapy**.

10.1.1 Classification of Drugs

Drugs can be classified mainly on criteria outlined as follows:

(a) On the basis of pharmacological effect

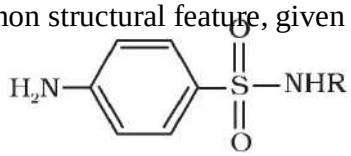
This classification is based on pharmacological effect of the drugs. It is useful for doctors because it provides them the whole range of drugs available for the treatment of a particular type of problem. For example, analgesics have pain killing effect, antiseptics kill or arrest the growth of microorganisms.

(b) On the basis of drug action

It is based on the action of a drug on a particular biochemical process. For example, all antihistamines inhibit the action of the compound, histamine which causes inflammation in the body. There are various ways in which action of histamines can be blocked. You will learn about this in Section 10.3.2.

(c) On the basis of chemical structure

It is based on the chemical structure of the drug. Drugs classified in this way share common structural features and often have similar pharmacological activity. For example, sulphonamides have common structural feature, given below.



Structural features of sulphonamides

(d) On the basis of molecular targets

Drugs usually interact with biomolecules such as carbohydrates, lipids, proteins and nucleic acids. These are called target molecules or drug targets. Drugs possessing some common structural features may have the same mechanism of action on targets. The classification based on molecular targets is the most useful classification for medicinal chemists.

10.2 Drug-Target Interaction

Macromolecules of biological origin perform various functions in the body. For example, proteins which perform the role of biological catalysts in the body are called **enzymes**, those which are crucial to communication system in the body are called **receptors**. Carrier proteins carry polar molecules across the cell membrane. Nucleic acids have coded genetic information for the cell. Lipids and carbohydrates are structural parts of the cell membrane. We shall explain the drug-target interaction with the examples of enzymes and receptors.

10.2.1 Enzymes as Drug Targets

(a) Catalytic action of enzymes

For understanding the interaction between a drug and an enzyme, it is important to know how enzymes catalyse the reaction (Section 5.2.4). In their catalytic activity, enzymes perform two major functions:

(i) The first function of an enzyme is to hold the substrate for a chemical reaction. Active sites of enzymes hold the substrate molecule in a suitable position, so that it can be attacked by the reagent effectively. Substrates bind to the active site of the enzyme through a variety of interactions such as ionic bonding, hydrogen bonding, van der Waals interaction or dipole-dipole interaction Fig. 10.1.

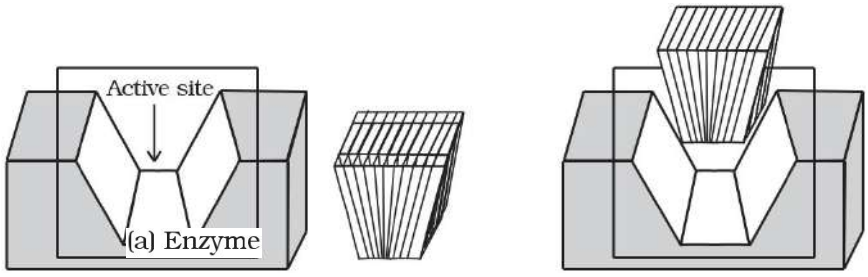


Fig. 10.1(a) Active site of an enzyme (b) Substrate
(c) Substrate held in active site of the enzyme

(ii) The second function of an enzyme is to provide functional groups that will attack the substrate and carry out chemical reaction.

(b) Drug-enzyme interaction

Drugs inhibit any of the above mentioned activities of enzymes. These can block the binding site of the enzyme and prevent the binding of substrate, or can inhibit the catalytic activity of the enzyme. Such drugs are called **enzyme inhibitors.**

Drugs inhibit the attachment of substrate on active site of enzymes in two different ways;

(i) Drugs compete with the natural substrate for their attachment on the active sites of enzymes. Such drugs are called **competitive inhibitors** Fig. 10.2.

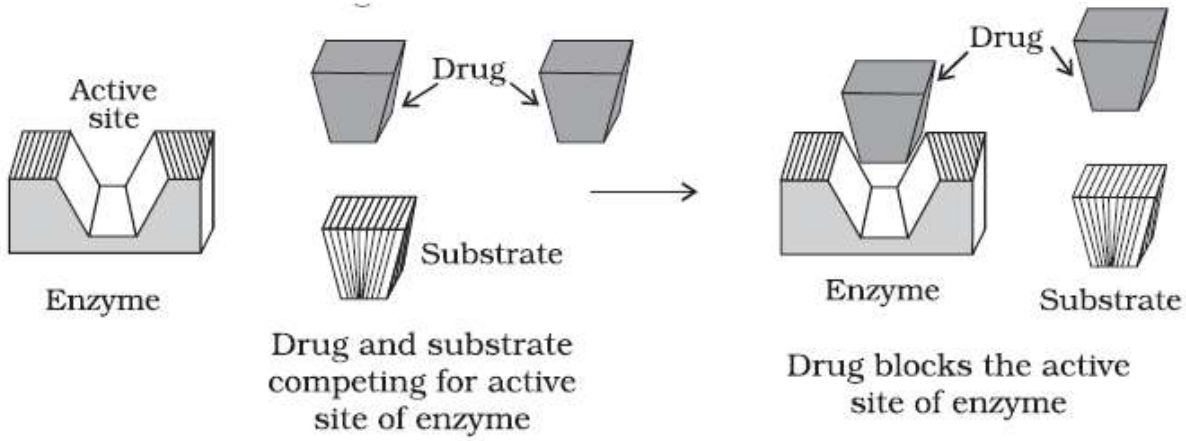


Fig. 10.2 Drug and substrate competing for active site

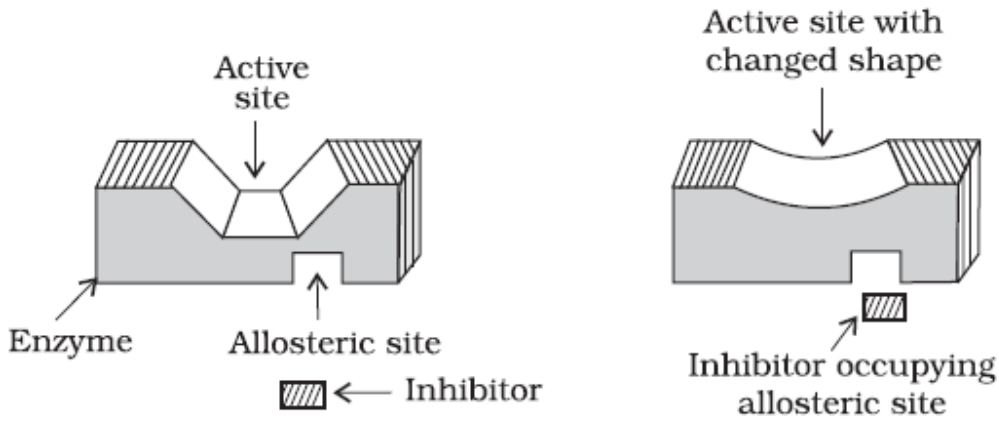


Fig. 10.3: Non-competitive inhibitor changes the active site of enzyme after binding at allosteric site.

(ii) Some drugs do not bind to the enzyme's active site. These bind to a different site of enzyme which is called **allosteric site**. This binding of inhibitor at allosteric site Fig.10.3 changes the shape of the active site in such a way that substrate cannot recognise it. If the bond formed between an enzyme and an inhibitor is a strong covalent bond and cannot be broken easily, then the enzyme is blocked permanently. The body then degrades the enzyme-inhibitor complex and synthesizes the new enzyme.

10.2.2 Receptors as Drug Targets

Receptors are proteins that are crucial to body's communication process. Majority of these are embedded in cell membranes (Fig.10.4). Receptor proteins are embedded in the cell membrane in such a way that their small part possessing active site projects out of the surface of the membrane and opens on the outside region of the cell membrane (Fig. 10.4).

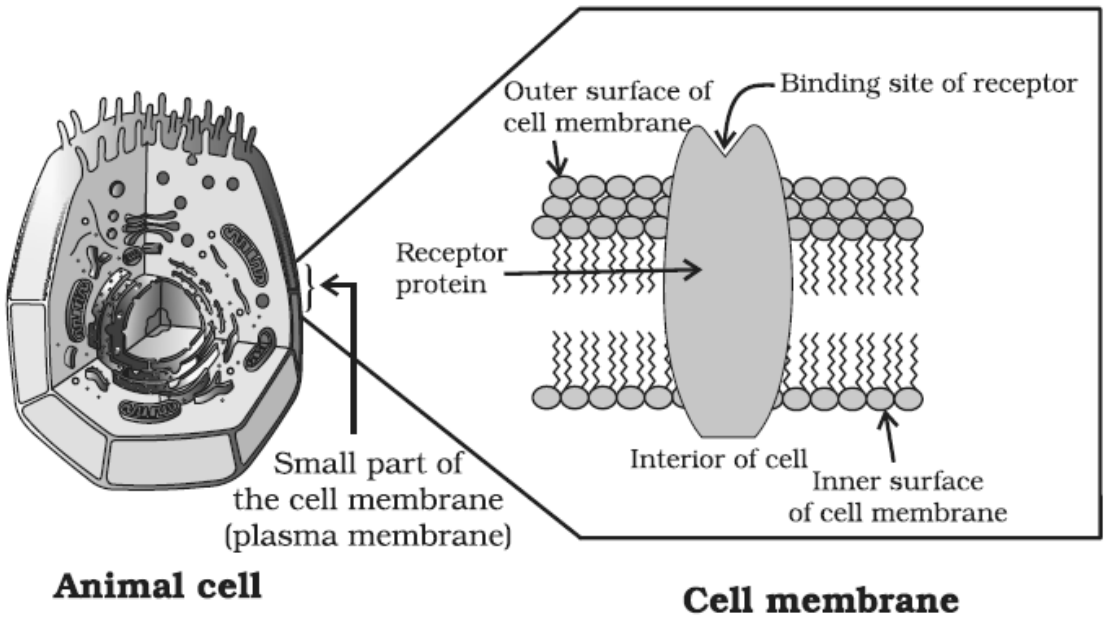


Fig. 10.4 Receptor protein embedded in the cell membrane, the active site of the receptor opens on the outside region of the cell.

In the body, message between two neurons and that between neurons to muscles is communicated through certain chemicals. These chemicals, known as **chemical messengers** are received at the binding sites of receptor proteins. To accommodate a messenger, shape of the receptor site changes. This brings about the transfer of message into the cell. Thus, chemical messenger gives message to the cell without entering the cell (Fig. 10.5).

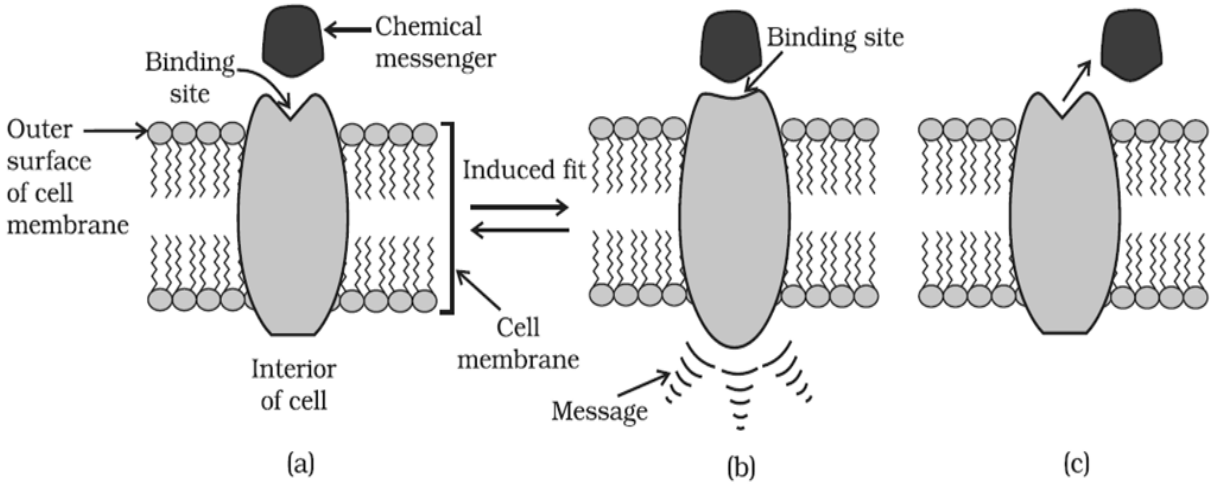


Fig. 10.5: (a) Receptor receiving chemical messenger
(b) Shape of the receptor changed after attachment of messenger
(c) Receptor regains structure after removal of chemical messenger.

There are a large number of different receptors in the body that interact with different chemical messengers. These receptors show selectivity for one chemical messenger over the other because their binding sites have different shape, structure and amino acid composition.

Drugs that bind to the receptor site and inhibit its natural function are called **antagonists**. These are useful when blocking of message is required. There are other types of drugs that mimic the natural messenger by switching on the receptor, these are called **agonists**. These are useful when there is lack of natural chemical messenger.

10.3 Therapeutic Action of Different Classes of Drugs

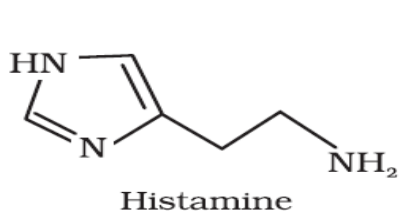
In this Section, we shall discuss the therapeutic action of a few important classes of drugs.

10.3.1 Antacids

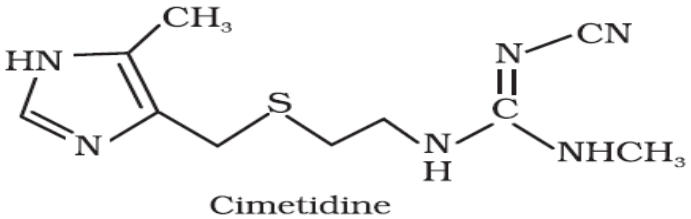
Over production of acid in the stomach causes irritation and pain. In severe cases, ulcers are developed in the stomach. Until 1970, only treatment for acidity was administration of antacids, such as sodium hydrogen carbonate or a mixture of aluminium and magnesium hydroxide. However, excessive hydrogen carbonate can make the stomach alkaline and trigger the production of even more acid. Metal hydroxides are better alternatives because of being insoluble; these do not increase the pH above neutrality. These treatments control only symptoms, and not the cause. Therefore, with these metal salts, the patients cannot be treated easily. In advanced stages, ulcers become life threatening and its only treatment is removal of the affected part of the stomach.

10.3.2 Antihistamines

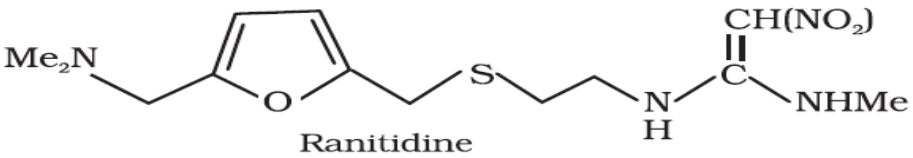
A major breakthrough in the treatment of hyperacidity came through the discovery according to which a chemical, histamine, stimulates the secretion of pepsin and hydrochloric acid in the stomach. The drug cimetidine (Tegamet), was designed to prevent the interaction of histamine with the receptors present in the stomach wall. This resulted in release of lesser amount of acid. The importance of the drug was so much that it remained the largest selling drug in the world until another drug, ranitidine (Zantac), was discovered.



Histamine

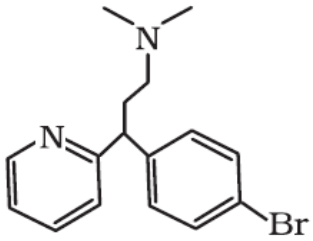


Cimetidine

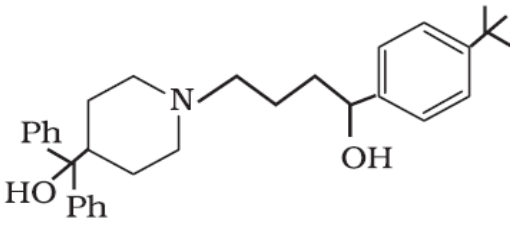


Ranitidine

Histamine is a potent vasodilator. It has various functions. It contracts the smooth muscles in the bronchi and gut and relaxes other muscles, such as those in the walls of fine blood vessels. Histamine is also responsible for the nasal congestion associated with common cold and allergic response to pollen.



Brompheniramine
(Dimetapp, Dimetane)



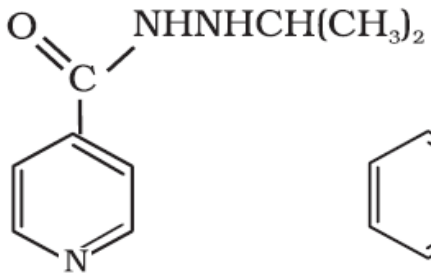
Terfenadine (Seldane)

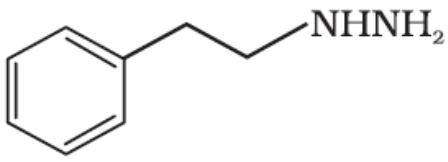
Synthetic drugs, **brompheniramine (Dimetapp)** and **terfenadine (Seldane),** act as antihistamines. They interfere with the natural action of histamine by competing with histamine for binding sites of receptor where histamine exerts its effect. Now the question that arises is, "Why do above mentioned antihistamines not affect the secretion of acid in stomach?" The reason is that anti-allergic and antacid drugs work on different receptors.

10.3.3 Neurologically Active Drugs

(a) Tranquilizers

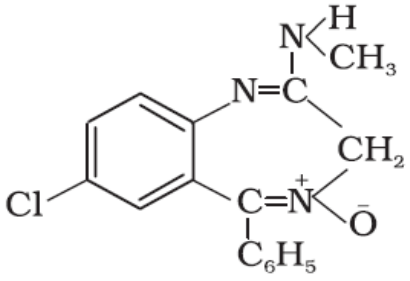
Tranquilizers and **analgesics** are neurologically active drugs. These affect the message transfer mechanism from nerve to receptor. Tranquilizers are a class of chemical compounds used for the treatment of stress, and mild or even severe mental diseases. These relieve anxiety, stress, irritability or excitement by inducing a sense of well-being. They form an essential component of sleeping pills. There are various types of tranquilizers. They function by different mechanisms. For example, noradrenaline is one of the neurotransmitters that plays a role in mood changes. If the level of noradrenaline is low for some reason, then the signal-sending activity becomes low, and the person suffers from depression. In such situations, **antidepressant drugs** are required. These drugs inhibit the enzymes which catalyse the degradation of noradrenaline. If the enzyme is inhibited, this important neurotransmitter is slowly metabolized and can activate its receptor for longer periods of time, thus counteracting the effect of depression. Iproniazid and Phenelzine are two such drugs.



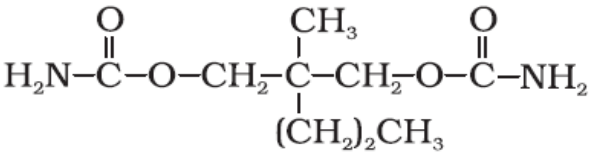


Iproniazid Phenelzine (Nardil)

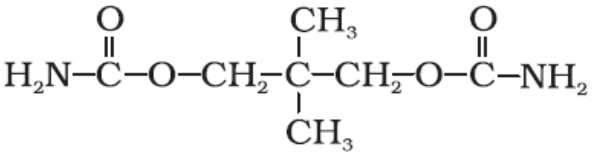
Some tranquilizers namely, chlordiazepoxide and meprobamate, are relatively mild tranquilizers suitable for relieving tension. Equanil is used in controlling depression and hypertension.



Chlordiazepoxide

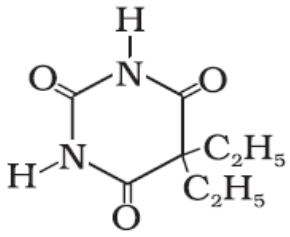


Meprobamate

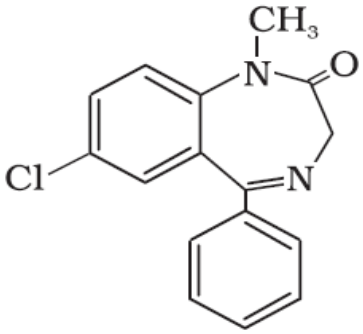


Equanil

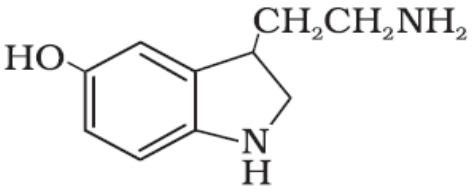
Derivatives of barbituric acid viz., veronal, amytal, nembutal, luminal and seconal constitute an important class of tranquilizers. These derivatives are called **barbiturates**. Barbiturates are hypnotic, *i.e.*, sleep producing agents. Some other substances used as tranquilizers are valium and serotonin.



Veronal



Valium



Serotonin

(b) Analgesics

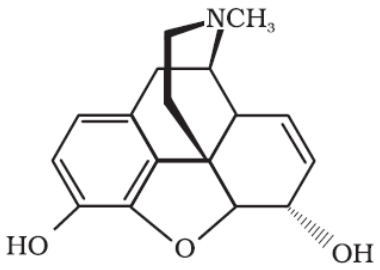
Analgesics reduce or abolish pain without causing impairment of consciousness, mental confusion, incoordination or paralysis or some other disturbances of nervous system. These are classified as follows:

(i) Non-narcotic (non-addictive) analgesics
(ii) Narcotic drugs

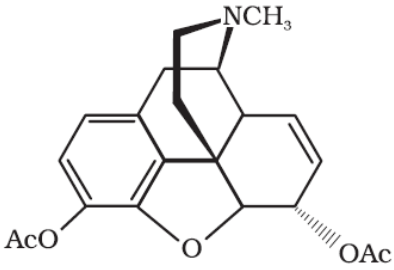
(i) ***Non-narcotic (non-addictive) analgesics:*** Aspirin and paracetamol belong to the class of **non-narcotic analgesics**. **Aspirin** is the most familiar example. Aspirin inhibits the synthesis of chemicals known as prostaglandins which stimulate inflammation in the tissue and cause pain. These drugs are effective in relieving skeletal pain such as that due to arthritis. These drugs have many other effects such as reducing fever **(antipyretic)** and preventing platelet coagulation. Because of its anti blood clotting action, aspirin finds use in prevention of heart attacks.

(ii) ***Narcotic analgesics:*** Morphine and many of its homologues, when administered in medicinal doses, relieve pain and produce sleep. In poisonous doses, these produce stupor, coma, convulsions and ultimately death. Morphine narcotics are sometimes referred to as opiates, since they are obtained from the opium poppy.

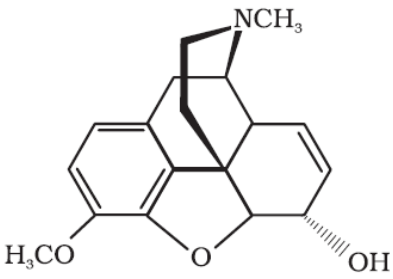
These analgesics are chiefly used for the relief of postoperative pain, cardiac pain and pains of terminal cancer, and in child birth.



Morphine



Heroin



Codeine

10.3.4 Antimicrobials

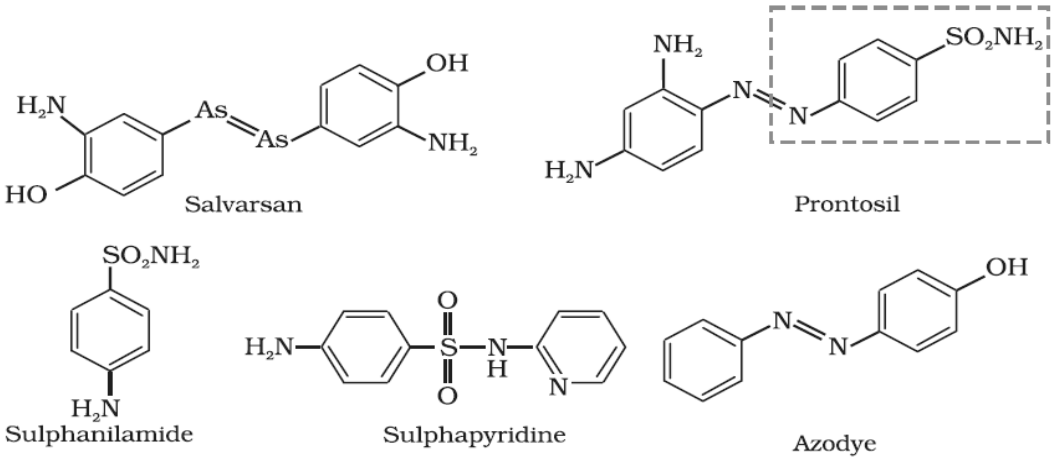
Diseases in human beings and animals may be caused by a variety of microorganisms such as bacteria, virus, fungi and other pathogens. An antimicrobial tends to destroy/prevent development or inhibit the pathogenic action of microbes such as bacteria (antibacterial drugs), fungi (antifungal agents), virus (antiviral agents), or other parasites (antiparasitic drugs) selectively. Antibiotics, antiseptics and disinfectants are antimicrobial drugs.

(a) Antibiotics

Antibiotics are used as drugs to treat infections because of their low toxicity for humans and animals. Initially antibiotics were classified as chemical substances produced by microorganisms (bacteria, fungi and molds) that inhibit the growth or even destroy microorganisms. The development of synthetic methods has helped in synthesizing some of the compounds that were originally discovered as products of microorganisms. Also, some purely synthetic compounds have antibacterial activity, and therefore, definition of antibiotic has been modified. An antibiotic now refers to a substance produced wholly or partly by chemical synthesis, which in low concentrations inhibits the growth or destroys microorganisms by intervening in their metabolic processes.

The search for chemicals that would adversely affect invading bacteria but not the host began in the nineteenth century. Paul Ehrlich, a German bacteriologist, conceived this idea. He investigated arsenic based structures in order to produce less toxic substances for the treatment of syphilis. He developed the medicine, **arsphenamine**, known as **salvarsan**. Paul Ehrlich got Nobel prize for Medicine in 1908 for this discovery. It was the first effective treatment discovered for syphilis. Although salvarsan is toxic to human beings, its effect on the bacteria, spirochete, which causes syphilis, is much greater than on human beings. At the same time, Ehrlich was working on azodyes also. He noted that there is similarity in structures of salvarsan and azodyes. The –As = As– linkage present in arsphenamine resembles the –N = N – linkage present in azodyes in the sense that arsenic atom is present in place of nitrogen. He also noted tissues getting coloured by dyes selectively. Therefore, Ehrlich began to search for the compounds which resemble in structure to azodyes and selectively bind to bacteria. In 1932, he succeeded in preparing the first effective antibacterial agent, **prontosil**, which resembles in structure to the compound, salvarsan. Soon it was discovered that in the body prontosil is converted to a compound called **sulphanilamide**, which is the real active compound. Thus the

sulpha drugs were discovered. A large range of sulphonamide analogues was synthesised. One of the most effective is sulphapyridine.



The structures of salvarsan, prontosil azodye and sulphapyridine showing structural similarity. H.W. Florey and Alexander Fleming shared the Nobel Prize for Medicine in 1945 for their independent contributions to the development of penicillin.

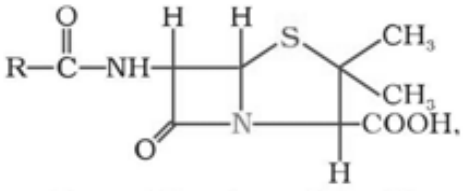
Despite the success of sulfonamides, the real revolution in antibacterial therapy began with the discovery of Alexander Fleming in 1929, of the antibacterial properties of a *Penicillium* fungus. Isolation and purification of active compound to accumulate sufficient material for clinical trials took thirteen years.

Antibiotics have either cidal (killing) effect or a static (inhibitory) effect on microbes. A few examples of the two types of antibiotics are as follows:

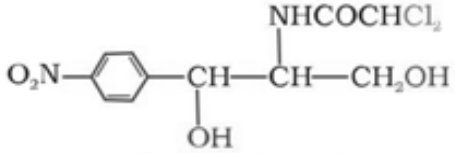
Bactericidal	**Bacteriostatic**
Penicillin	Erythromycin
Aminoglycosides	Tetracycline
Ofloxacin	Chloramphenicol

The range of bacteria or other microorganisms that are affected by a certain antibiotic is expressed as its spectrum of action. Antibiotics which kill or inhibit a wide range of Gram-positive and Gram-negative bacteria are said to be **broad spectrum antibiotics**. Those effective mainly against Gram-positive or Gram-negative bacteria are **narrow spectrum antibiotics**. If effective against a single organism or disease, they are referred to as **limited spectrum** antibiotics. Penicillin G has a narrow spectrum. Ampicillin and Amoxycillin are synthetic modifications of penicillins. These have broad spectrum. It is absolutely essential to test the patients for sensitivity (allergy) to penicillin before it is administered. In India, penicillin is manufactured at the Hindustan Antibiotics in Pimpri and in private sector industry.

Chloramphenicol, isolated in 1947, is a broad spectrum antibiotic. It is rapidly absorbed from the gastrointestinal tract and hence can be given orally in case of typhoid, dysentery, acute fever, certain form of urinary infections, meningitis and pneumonia. *Vancomycin* and *ofloxacin* are the other important broad spectrum antibiotics. The antibiotic *dysidazirine* is supposed to be toxic towards certain strains of cancer cells.



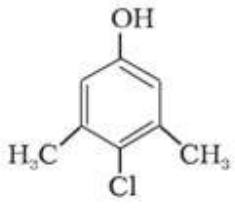
General Sturcture of Pencillin



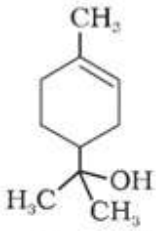
Chloramphenicol

(b) Antiseptics and disinfectants

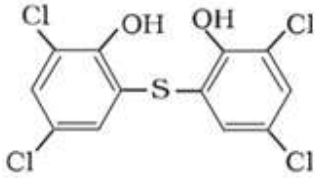
Antiseptics and disinfectants are also the chemicals which either kill or prevent the growth of microorganisms. **Antiseptics** are applied to the living tissues such as wounds, cuts, ulcers and diseased skin surfaces. Examples are **furacine**, **soframicine,** etc. These are not ingested like antibiotics. Commonly used antiseptic, dettol is a mixture of **chloroxylenol** and **terpineol**. Bithionol (the compound is also called bithional) is added to soaps to impart antiseptic properties.



Chloroxylenol



Terpineol



Bithionol

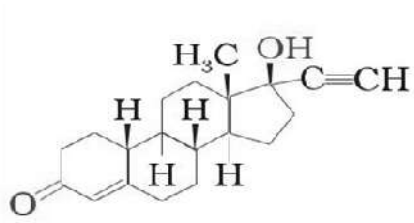
Iodine is a powerful antiseptic. Its 2-3 per cent solution in alcohol-water mixture is known as **tincture of iodine**. It is applied on wounds. **Iodoform** is also used as an antiseptic for wounds. Boric acid in dilute aqueous solution is weak antiseptic for eyes.

Disinfectants are applied to inanimate objects such as floors, drainage system, instruments, etc. Same substances can act as an antiseptic as well as disinfectant by varying the concentration. For example, 0.2 per cent solution of phenol is an antiseptic while its one percent solution is disinfectant.

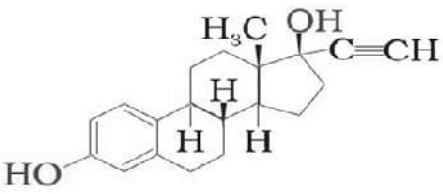
Chlorine in the concentration of 0.2 to 0.4 ppm in aqueous solution and sulphur dioxide in very low concentrations, are disinfectants.

10.3.5 Anti-fertility Drugs

Antibiotic revolution has provided long and healthy life to people. The life expectancy has almost doubled. The increased population has caused many social problems in terms of food resources, environmental issues, employment, etc. To control these problems, population is required to be controlled. This has lead to the concept of family planning. Anti-fertility drugs are of use in this direction. Birth control pills essentially contain a mixture of synthetic estrogen and progesterone derivatives. Both of these compounds are hormones. It is known that progesterone suppresses ovulation. Synthetic progesterone derivatives are more potent than progesterone. **Norethindrone** is an example of synthetic progesterone derivative most widely used as antifertility drug. The estrogen derivative which is used in combination with progesterone derivative is **ethynylestradiol (novestrol)**.



Norethindrone



Ethynylestradiol (novestrol)

10 .4 Chemicals in Food

Chemicals are added to food for (i) their preservation, (ii) enhancing their appeal, and (iii) adding nutritive value in them. Main categories of food additives are as follows:

(i) Food colours
(ii) Flavours and sweeteners
(iii) Fat emulsifiers and stabilizing agents
(iv) Flour improvers - antistaling agents and bleaches
(v) Antioxidants
(vi) Preservatives
(vii) Nutritional supplements such as minerals, vitamins and amino acids.

Except for chemicals of category (vii), none of the above additives have nutritive value. These are added either to increase the shelf life of stored food or for cosmetic purposes. In this Section we will discuss only sweeteners and food preservatives.

10.4.1 Artificial Sweetening Agents

Natural sweeteners, e.g., sucrose add to calorie intake and therefore many people prefer to use artificial sweeteners. Ortho-sulphobenzimide, also called saccharin, is the first popular artificial sweetening agent. It has been used as a sweetening agent ever since it was discovered in 1879. It is about 550 times as sweet as cane sugar. It is excreted from the body in urine unchanged. It appears to be entirely inert and harmless when taken. Its use is of great value to diabetic persons and people who need to control intake of calories. Some other commonly marketed artificial sweeteners are given in Table 10.1.

Aspartame is the most successful and widely used artificial sweetener. It is roughly 100 times as sweet as cane sugar. It is methyl ester of dipeptide formed from aspartic acid and phenylalanine. Use of aspartame is limited to cold foods and soft drinks because it is unstable at cooking temperature.

Alitame is high potency sweetener, although it is more stable than aspartame, the control of sweetness of food is difficult while using it.

Sucrolose is trichloro derivative of sucrose. Its appearance and taste are like sugar. It is stable at cooking temperature. It does not provide calories.

10.4.2 Food Preservatives

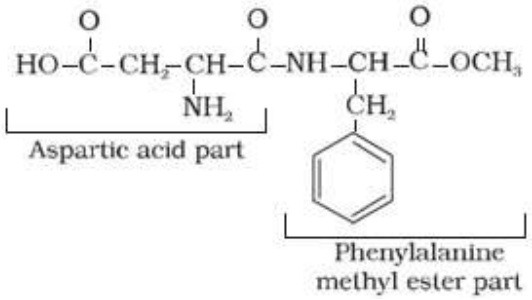
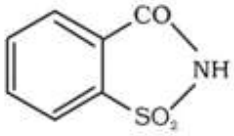
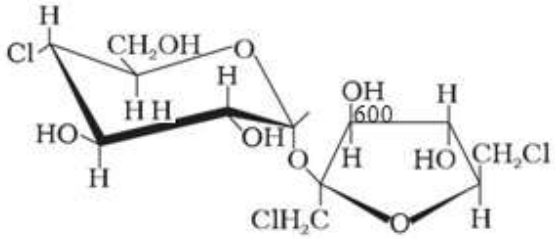
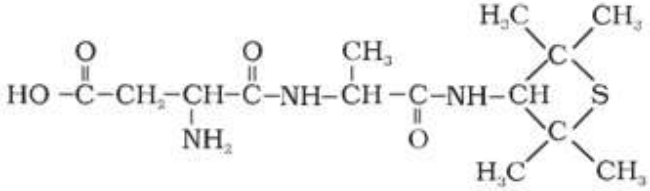
Food preservatives prevent spoilage of food due to microbial growth. **Preservatives** The most commonly used preservatives include table salt, sugar, vegetable oils and sodium benzoate,

C_6H_5COONa. Sodium benzoate is used in limited quantities and is metabolised in the body. Salts of sorbic acid and propanoic acid are also used as preservatives.

Summary

Chemistry is essentially the study of materials and the development of new materials for the betterment of humanity. A **drug** is a chemical agent, which affects human metabolism and provides cure from ailment. If taken in doses higher than recommended, these may have poisonous effect. Use of chemicals for therapeutic effect is called **chemotherapy**. Drugs usually interact with biological macromolecules such as carbohydrates, proteins, lipids and nucleic acids. These are called **target molecules**.

Table 10.1: Artificial Sweeteners

Artificial sweetener	Structural formula	Sweetness value in comparison to cane sugar
Co		
Aspartame		**100**
Saccharin		**550**
Sucrolose		**600**
Alitame		**2000**

Drugs are designed to interact with specific targets so that these have the least chance of affecting other targets. This minimises the side effects and localises the action of the drug. Drug chemistry centres around arresting microbes/destroying microbes, preventing the body from

various infectious diseases, releasing mental stress, etc. Thus, drugs like analgesics, antibiotics, antiseptics, disinfectants, antacids and tranquilizers are used for specific purpose. To check the population explosion, anti-fertility drugs have also become prominent in our life.

Food additives such as **preservatives**, **sweetening agents**, **flavours**, **antioxidants, edible colours** and **nutritional supplements** are added to the food to make it attractive, palatable and add nutritive value. Preservatives are added to the food to prevent spoilage due to microbial growth. Artificial sweeteners are used by those who need to check the calorie intake or are diabetic and want to avoid taking sucrose. These days, **detergents** are much in vogue and get preference over soaps because they work even in hard water. Synthetic detergents are classified into three main categories, namely: **anionic, cationic** and **non-ionic,** and each category has its specific uses. Detergents with straight chain of hydrocarbons are preferred over branched chain as the latter are **non-biodegradable** and consequently cause **environmental pollution**.

IMPORTANT QUESTIONS

1. What are Analgesics? How are they Classified? Give Examples
2. Define
   (i) Antimicrobials
   (ii) Antibiotics
   (iii) Antiseptis
   (iv) Disinfectants
   (v) Artificial Sweetening Agents
   (vi) Food Preservatives
3. What is Tincture of Iodine? What is its use?
4. What are antacids? Give one Example
5. What are Antihistamines. Give one example

Chapter 11

HALOALKANES & HALOARENES

The replacement of hydrogen atom(s) in a hydrocarbon, aliphatic or aromatic, by halogen atom(s) results in the formation of alkyl halide (haloalkane) and aryl halide (haloarene), respectively. Haloalkanes contain halogen atom(s) attached to the sp^3 hybridised carbon atom of an alkyl group whereas haloarenes contain halogen atom(s) attached to sp^2 hybridised carbon atom(s) of an aryl group. Many halogen containing organic compounds occur in nature and some of these are clinically useful. These classes of compounds find wide applications in industry as well as in day-to-day life. They are used as solvents for relatively non-polar compounds and as starting materials for the synthesis of wide range of organic compounds. Chlorine containing antibiotic, *chloramphenicol,* produced by soil microorganisms is very effective for the treatment of typhoid fever. Our body produces iodine containing hormone, *thyroxine*, the deficiency of which causes a disease called *goiter*. Synthetic halogen compounds, *viz.* chloroquine is used for the treatment of malaria; halothane is used as an anaesthetic during surgery. Certain fully fluorinated compounds are being considered as potential blood substitutes in surgery.

In this Unit, you will study the important methods of preparation, physical and chemical properties and uses of organohalogen compounds.

11.1 Classification

Haloalkanes and haloarenes may be classified as follows

On the Basis of Number of Halogen Atoms

These may be classified as mono, di, or polyhalogen (tri-,tetra-, etc.) compounds depending on whether they contain one, two or more halogen atoms in their structures. For example,

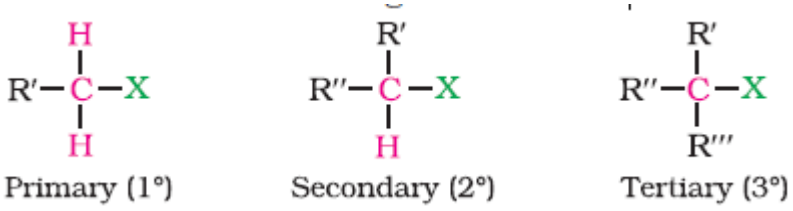
Mono halo compounds may further be classified according to the hybridisation of the carbon atom to which the halogen is bonded, as discussed below.

Compounds Containing *sp3* C—X Bond (X= F, Cl, Br, I)

This class includes

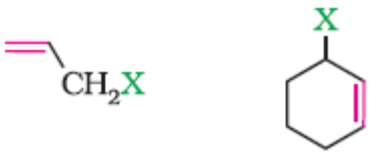
(a) Alkyl halides or haloalkanes (R—X)

In alkyl halides, the halogen atom is bonded to an alkyl group (R). They form a homologous series represented by $C_nH_{2n+1}X$. They are further classified as primary, secondary or tertiary according to the nature of carbon to which halogen is attached.



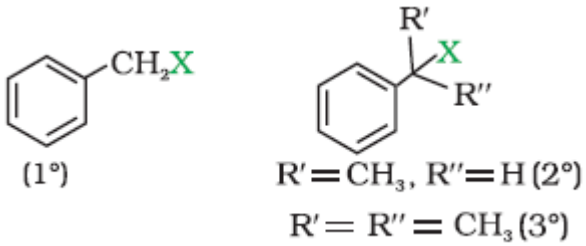
(b) Allylic halides

These are the compounds in which the halogen atom is bonded to an *sp3*-hybridised carbon atom next to carbon-carbon double bond (C=C) *i.e.* to an allylic carbon.



(c) Benzylic halides

These are the compounds in which the halogen atom is bonded to an *sp3*-hybridised carbon atom next to an aromatic ring.



Compounds Containing *sp2* C—X Bond

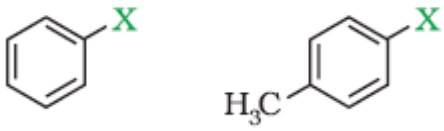
This class includes:

(a) Vinylic halides

These are the compounds in which the halogen atom is bonded to an *sp2*-hybridised carbon atom of a carbon-carbon double bond (C = C).

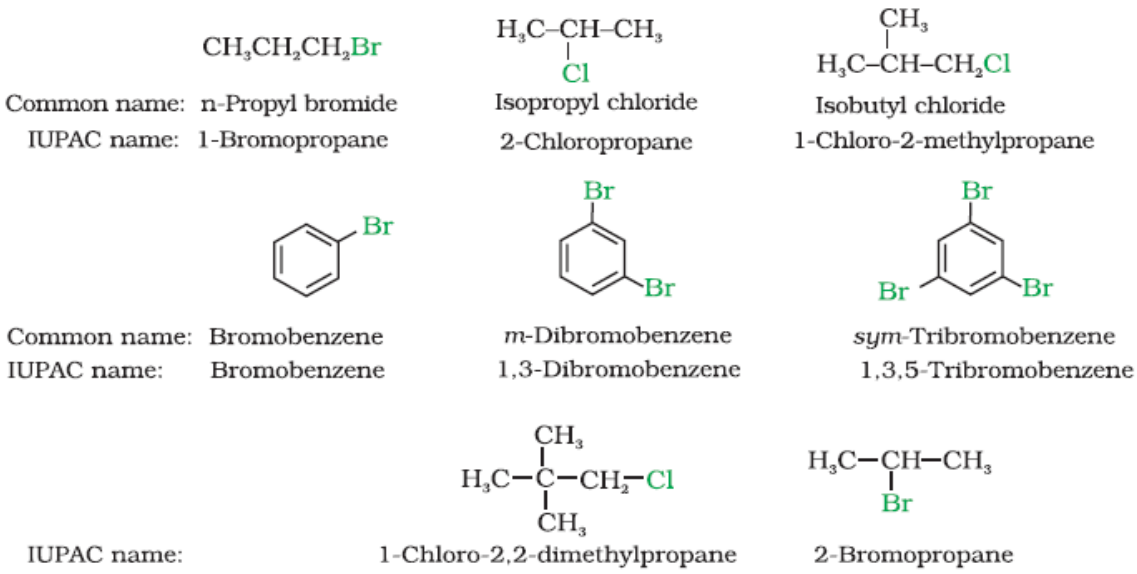
b) Aryl halides

These are the compounds in which the halogen atom is bonded to the sp2-hybridised carbon atom of an aromatic ring.



11.2 Nomenclature

Having learnt the classification of halogenated compounds, let us now learn how these are named. The common names of alkyl halides are derived by naming the alkyl group followed by the halide. Alkyl halides are named as halosubstituted hydrocarbons in the IUPAC system of nomenclature. Haloarenes are the common as well as IUPAC names of aryl halides. For dihalogen derivatives, the prefixes *o*-, *m*-, *p*- are used in common system but in IUPAC system, the numerals 1,2; 1,3 and 1,4 are used.



The dihaloalkanes having the same type of halogen atoms are named as alkylidene or alkylene dihalides. The dihalo-compounds having same type of halogen atoms are further classified as geminal halides (halogen atoms are present on the same carbon atom) and vicinal halides (halogen atoms are present on the adjacent carbon atoms). In common name system, *gem*-dihalides are named as alkylidene halides and *vic*-dihalides are named as alkylene dihalides. In IUPAC system, they are named as dihaloalkanes.

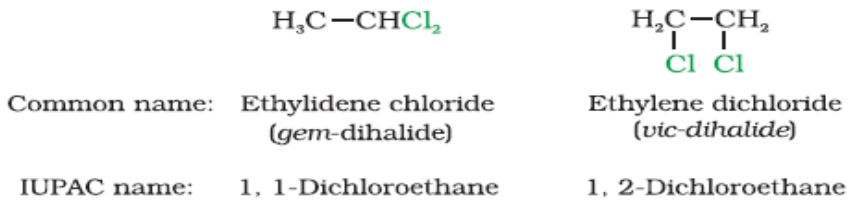
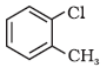
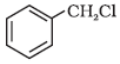
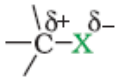


Table : Common and IUPAC names of some Halides

Structure	Common name	IUPAC name
$CH_3CH_2CH(Cl)CH_3$	sec-Butyl chloride	2-Chlorobutane
$(CH_3)_3CCH_2Br$	neo-Pentyl bromide	1-Bromo-2,2-dimethylpropane
$(CH_3)_3CBr$	tert-Butyl bromide	2-Bromo-2-methylpropane
$CH_2 = CHCl$	Vinyl chloride	Chloroethene
$CH_2 = CHCH_2Br$	Allyl bromide	3-Bromopropene
	*o*-Chlorotoluene	1-Chloro-2-methylbenzene Or 2-Chlorotoluene
	Benzyl chloride	Chlorophenylmethane
CH_2Cl_2	Methylene chloride	Dichloromethane
$CHCl_3$	Chloroform	Trichloromethane
$CHBr_3$	Bromoform	Tribromomethane
CCl_4	Carbon tetrachloride	Tetrachloromethane
$CH_3CH_2CH_2F$	n-Propyl fluoride	1-Fluoropropane

11.3 Nature of C-X Bond

Since halogen atoms are more electronegative than carbon, the carbonhalogen bond of alkyl halide is polarised; the carbon atom bears a partial positive charge whereas the halogen atom bears a partial negative charge.



Since the size of halogen atom increases as we go down the group in the periodic table, fluorine atom is the smallest and iodine atom, the largest. Consequently the carbon-halogen bond length also increases from C—F to C—I. Some typical bond lengths, bond enthalpies and dipole moments are given in Table 10.2.

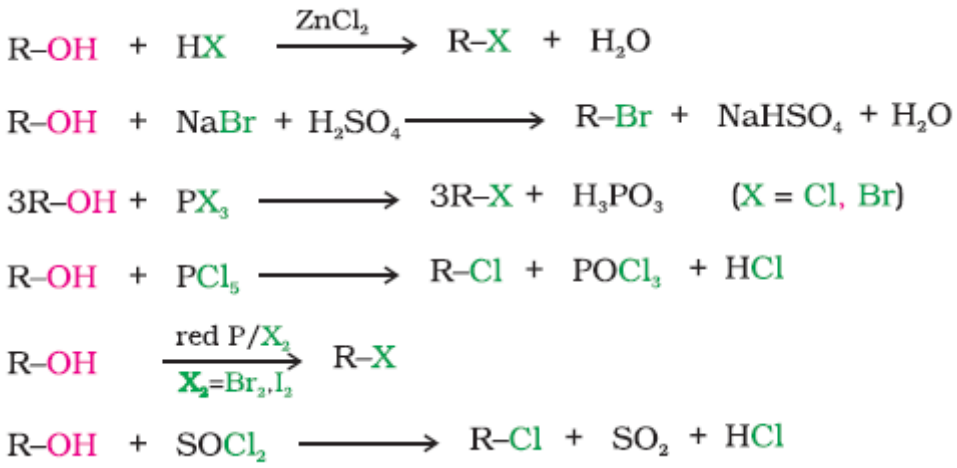
Table : Carbon-Halogen (C—X) Bond Lengths, Bond Enthalpies and Dipole Moments

Bond	Bond length/pm	C-X Bond enthalpies/ kJmol^{-1}	Dipole moment/Debye
CH_3–F	139	452	1.847
CH_3– Cl	178	351	1.860
CH_3–Br	193	293	1.830
CH_3–I	214	234	1.636

11.4 Methods of Preparation

11.4.1 From Alcohols

Alkyl halides are best prepared from alcohols, which are easily accessible. The hydroxyl group of an alcohol is replaced by halogen on reaction with concentrated halogen acids, phosphorus halides or thionyl chloride Thionyl chloride is preferred because the other two products are escapable gases. Hence the reaction gives pure alkyl halides. Phosphorus tribromide and triiodide are usually generated *in situ* (produced in the reaction mixture) by the reaction of red phosphorus with bromine and iodine respectively. The preparation of alkyl chloride is carried out either by passing dry hydrogen chloride gas through a solution of alcohol or by heating a solution of alcohol in concentrated aqueous acid.



The reactions of primary and secondary alcohols with HX require the presence of a catalyst, ZnCl2. With tertiary alcohols, the reaction is conducted by simply shaking with concentrated HCl at room temperature. Constant boiling with HBr (48%) is used for

preparing alkyl bromide. Good yields of R—I may be obtained by heating alcohols with sodium or potassium iodide in 95% phosphoric acid. The order of reactivity of alcohols with a given haloacid is 3°>2°>1°. The above method is not applicable for the preparation of aryl halides because the carbon-oxygen bond in phenols has a partial double bond character and is difficult to break being stronger than a single bond.

From Hydrocarbons

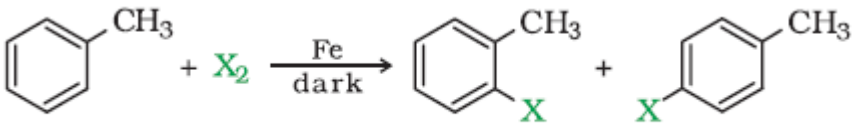
(a) By free radical halogenations

Free radical chlorination or bromination of alkanes gives a complex mixture of isomeric mono- and polyhaloalkanes, which is difficult to separate as pure compounds. Consequently, the yield of any one compound is low.

$$CH_3CH_2CH_2CH_3 \xrightarrow[\text{or heat}]{Cl_2/\text{UV light}} CH_3CH_2CH_2CH_2Cl + CH_3CH_2CHClCH_3$$

(b) By electrophilic substitution

Aryl chlorides and bromides can be easily prepared by electrophilic substitution of arenes with chlorine and bromine respectively in the presence of Lewis acid catalysts like iron or iron(III) chloride.



o-Halotoluene *p*-Halotoluene

The *ortho* and *para* isomers can be easily separated due to large difference in their melting points. Reactions with iodine are reversible in nature and require the presence of an oxidising agent (HNO3, HIO4) to oxidise the HI formed during iodination. Fluoro compounds are not prepared by this method due to high reactivity of fluorine.

11.5 Physical Properties

Alkyl halides are colourless when pure. However, bromides and iodides develop colour when exposed to light. Many volatile halogen compounds have sweet smell.

Melting and boiling points

Methyl chloride, methyl bromide, ethyl chloride and some chlorofluoromethanes are gases at room temperature. Higher members are liquids or solids. As we have already learnt, molecules of organic halogen compounds are generally polar. Due to greater polarity as well as higher molecular mass as compared to the parent hydrocarbon, the intermolecular forces of attraction (dipole-dipole and van der Waals) are stronger in the halogen derivatives. That is why the boiling points of chlorides, bromides and iodides are considerably higher than those of the hydrocarbons of comparable molecular mass.

The attractions get stronger as the molecules get bigger in size and have more electrons. The pattern of variation of boiling points of different halides is depicted in Fig. 10.1. For the same alkyl group, the boiling points of alkyl halides decrease in the order: RI> RBr> RCl> RF. This is because with the increase in size and mass of halogen atom, the magnitude of van der Waal forces increases.

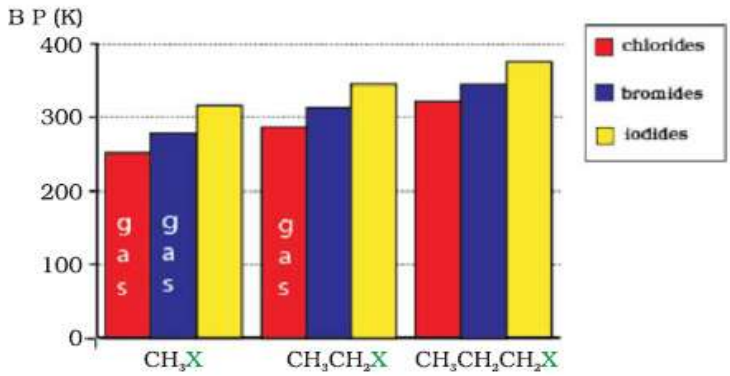


Fig. 11.1: Comparison of boiling points of some alkyl halides

The boiling points of isomeric haloalkanes decrease with increase in branching (Unit 12, Class XI). For example, 2-bromo-2-methylpropane has the lowest boiling point among the three isomers.

	$CH_3CH_2CH_2CH_2Br$	$CH_3CH_2CHCH_3$ (Br)	$H_3C-C(CH_3)_2$ (Br)
b.p./K	375	364	346

Density

Bromo, iodo and polychloro derivatives of hydrocarbons are heavier than water. The density increases with increase in number of carbon atoms, halogen atoms and atomic mass of the halogen atoms.

Solubility

The haloalkanes are only very slightly soluble in water. In order for a haloalkane to dissolve in water, energy is required to overcome the attractions between the haloalkane molecules and break the hydrogen bonds between water molecules. Less energy is released when new attractions are set up between the haloalkane and the water molecules as these are not as strong as the original hydrogen bonds in water. As a result, the solubility of aloalkanes in water is low. However, haloalkanes tend to dissolve in organic solvents because the new intermolecular attractions between haloalkanes and solvent molecules have much the same strength as the ones being broken in the separate haloalkane and solvent molecules.

11.6 Chemical Reactions

Reactions of Haloalkanes

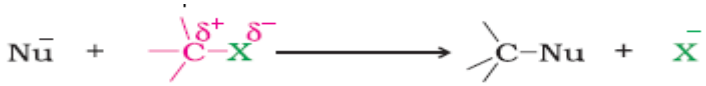
The reactions of haloalkanes may be divided into the following categories:

(i) Nucleophilic substitution

(ii) Elimination reactions

(i) *Nucleophilic substitution reactions*

In this type of reaction, a nucleophile reacts with haloalkane (the substrate) having a partial positive charge on the carbon atom bonded to halogen. A substitution reaction takes place and halogen atom, called leaving group departs as halide ion. Since the substitution reaction is initiated by a nucleophile, it is called nucleophilic substitution reaction.

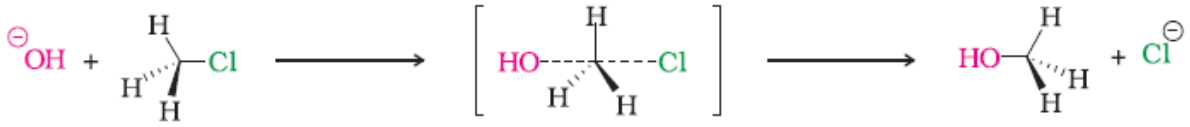


Mechanism

This reaction has been found to proceed by two different mechanims which are described below:

(a) Substitution nucleophilic bimolecular (SN2)

The reaction between CH_3Cl and hydroxide ion to yield methanol and chloride ion follows a second order kinetics, i.e., the rate depends upon the concentration of both the reactants.



It depicts a bimolecular nucleophilic displacement (SN2) reaction; the incoming nucleophile interacts with alkyl halide causing the carbonhalide bond to break while forming a new carbon-OH bond. These two processes take place simultaneously in a single step and no intermediate is formed. As the reaction progresses and the bond between the nucleophile and the carbon atom starts forming, the bond between carbon atom and leaving group weakens. As this happens, the configuration of carbon atom under attack inverts in much the same way as an umbrella is turned inside out when caught in a strong wind, while the leaving group is pushed away. This process is called as **inversion of configuration**. In the transition state, the carbon atom is simultaneously bonded to incoming nucleophile and the outgoing leaving group and such structures are unstable and cannot be isolated. This is because the carbon atom in the transition state is simultaneously bonded to five atoms and therefore is unstable.

Since this reaction requires the approach of the nucleophile to the carbon bearing the leaving group, the presence of bulky substituents on or near the carbon atom have a dramatic inhibiting effect. Of the simple alkyl halides, methyl halides react most rapidly in SN2 reactions because there are only three small hydrogen atoms. Tertiary halides are the least reactive because bulky groups hinder the approaching nucleophiles. Thus the order of reactivity followed is:

Primary halide > Secondary halide > Tertiary halide.

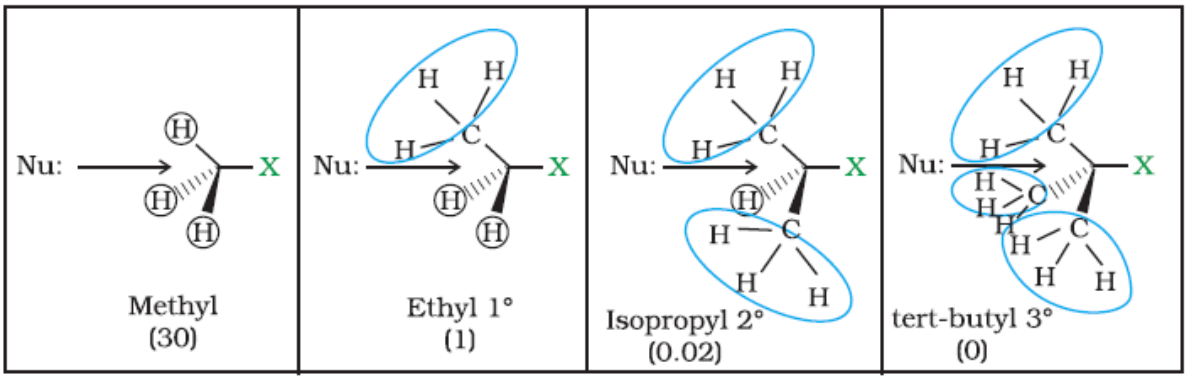


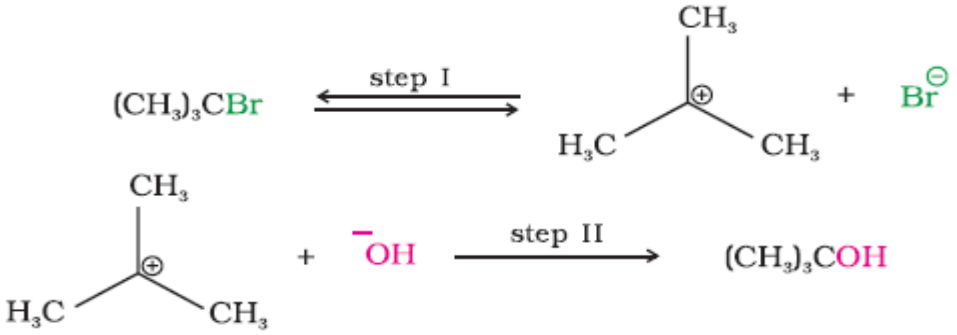
Fig.11.3 Steric effects in SN^2 reaction. The relative rate of SN^2 reaction is given in parenthesis

(b) Substitution nucleophilic unimolecular (SN1)

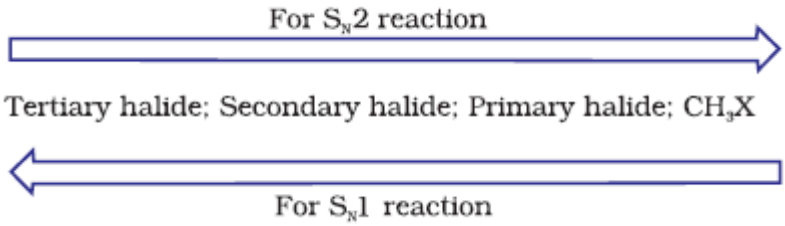
SN1 reactions are generally carried out in polar protic solvents (like water, alcohol, acetic acid, etc.). The reaction between *tert*-butyl bromide and hydroxide ion yields *tert*-butyl alcohol and follows the first order kinetics, *i.e.*, the rate of reaction depends upon the concentration of only one reactant, which is *tert*- butyl bromide.

1

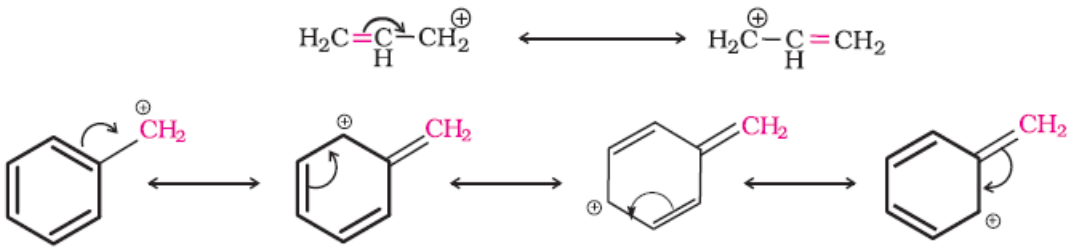
It occurs in two steps. In step I, the polarised C—Br bond undergoes slow cleavage to produce a carbocation and a bromide ion. The carbocation thus formed is then attacked by nucleophile in step II to complete the substitution reaction.



Step I is the slowest and reversible. It involves the C–Br bond breaking for which the energy is obtained through solvation of halide ion with the proton of protic solvent. Since the rate of reaction depends upon the slowest step, the rate of reaction depends only on the concentration of alkyl halide and not on the concentration of hydroxide ion. Further, greater the stability of carbocation, greater will be its ease of formation from alkyl halide and faster will be the rate of reaction. In case of alkyl halides, 30 alkyl halides undergo S_N1 reaction very fast because of the high stability of 30 carbocations. We can sum up the order of reactivity of alkyl halides towards S_N1 and S_N2 reactions as follows:



For the same reasons, allylic and benzylic halides show high reactivity towards the S_N1 reaction. The carbocation thus formed gets stabilised through resonance as shown below.



For a given alkyl group, the reactivity of the halide, R-X, follows the same order in both the mechanisms R–I> R–Br>R–Cl>>R–F.

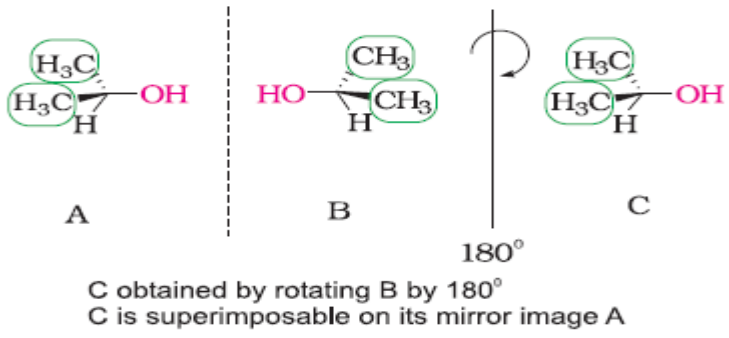
(c) Stereochemical aspects of nucleophilic substitution reactions

A SN2 reaction proceeds with complete stereochemical inversion while a SN1 reaction proceeds with racemisation. In order to understand this concept, we need to learn some basic stereochemical principles and notations (**optical activity, chirality, retention, inversion, racemisation,** etc.).

(i) Plane polarised light and optical activity: Certain compounds rotate the plane polarised light (produced by passing ordinary light through Nicol prism) when it is passed through their solutions. Such compounds are called **optically active** compounds. The angle by which the plane polarised light is rotated is measured by an instrument called polarimeter. If the compound rotates the plane polarised light to the right, i.e., clockwise direction, it is called dextrorotatory (Greek for right rotating) or the d-form and is indicated by placing a positive (+) sign before the degree of rotation. If the light is rotated towards left (anticlockwise direction), the compound is said to be laevorotatory or the l-form and a negative (–) sign is placed before the degree of rotation. Such (+) and (–) isomers of a compound are called **optical isomers** and the phenomenon is termed as **optical isomerism**.

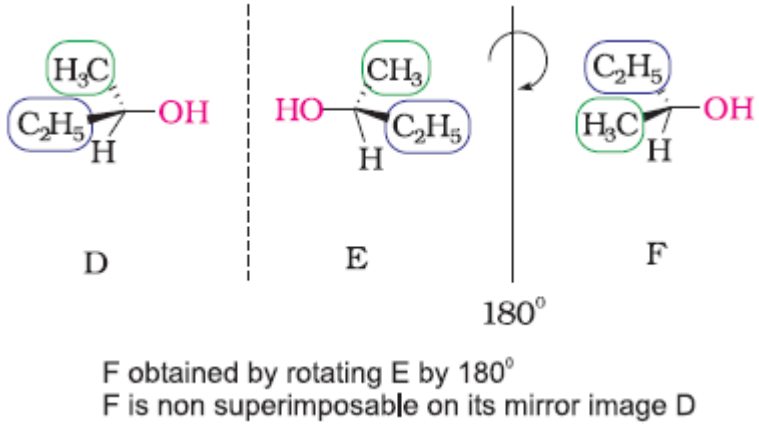
(ii) Molecular asymmetry, chirality and enantiomers: The observation of Louis Pasteur (1848) that crystals of certain compounds exist in the form of mirror images laid the foundation of modern stereochemistry. He demonstrated that aqueous solutions of both types of crystals showed optical rotation, equal in magnitude (for solution of equal concentration) but opposite in direction. He believed that this difference in optical activity was associated with the three dimensional arrangements of atoms (**configurations**) in two types of crystals. Dutch scientist, *J. Van't Hoff* and French scientist, *C. Le Bel* in the same year (1874), independently argued that the spatial arrangement of four groups (valencies) around a central carbon is tetrahedral and if all the substituents attached to that carbon are different, such a carbon is called **asymmetric carbon** or **stereocentre**. The resulting molecule would lack symmetry and is referred to as asymmetric molecule. The asymmetry of the molecule is responsible for the optical activity in such organic compounds.

The molecular chirality can be applied to organic molecules by constructing models and its mirror images or by drawing three dimensional structures and attempting to superimpose them in our minds. There are other aids, however, that can assist us in recognising chiral molecules. One such aid is the presence of a single asymmetric carbon atom. Let us consider two simple molecules propan-2-ol and butan-2-ol and their mirror images.



C obtained by rotating B by 180°
C is superimposable on its mirror image A

As you can see very clearly, propan-2-ol does not contain an asymmetric carbon, as all the four groups attached to the tetrahedral carbon are not different. Thus it is an **achiral** molecule.



F obtained by rotating E by 180°
F is non superimposable on its mirror image D

Butan-2-ol has four different groups attached to the tetrahedral carbon and as expected is **chiral**. Some common examples of chiral molecules such as 2-chlorobutane, 2, 3-dihyroxypropanal,(OHC–CHOH–CH_2OH), bromochloro-iodomethane (BrClCHI), 2-bromopropanoic acid (H_3C–CHBr–COOH), etc.

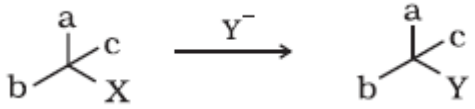
The stereoisomers related to each other as nonsuperimposable mirror images are called **enantiomers**

Enantiomers possess identical physical properties namely, melting point, boiling point, solubility, refractive index, etc. They only differ with respect to the rotation of plane polarised light. If one of the enantiomer is *dextro rotatory*, the other will be *laevo rotatory*.

A mixture containing two enantiomers in equal proportions will have zero optical rotation, as the rotation due to one isomer will be cancelled by the rotation due to the other isomer. Such a mixture is known as **racemic mixture** or **racemic modification**. A racemic mixture is represented by prefixing *dl* or (±) before the name, for example (±) butan-2-ol. The process of conversion of enantiomer into a racemic mixture is known as **racemisation.**

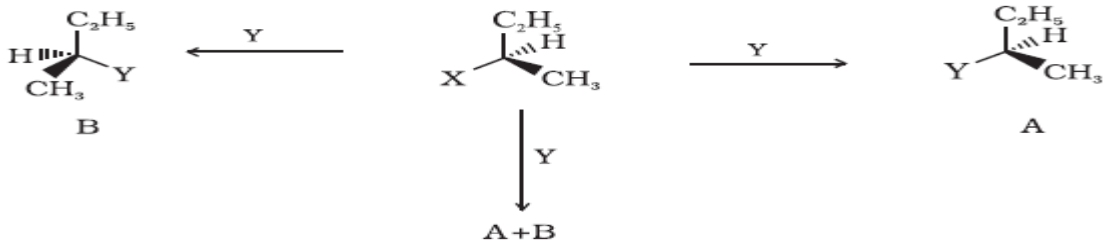
(iii) **Retention:**

Retention of configuration is the preservation of integrity of the spatial arrangement of bonds to an asymmetric centre during a chemical reaction or transformation. It is also the configurational correlation when a chemical species XCabc is converted into the chemical species YCabc having the same *relative configuration*.



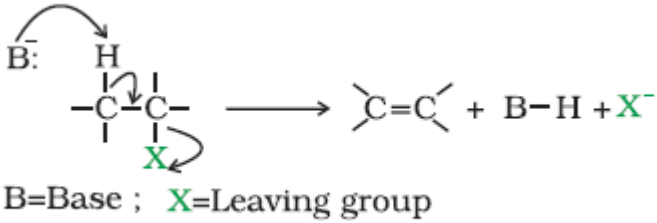
(iv) **Inversion, retention and racemisation**:

There are three outcomes for a reaction at an asymmetric carbon atom. Consider the replacement of a group X by Y in the following reaction;

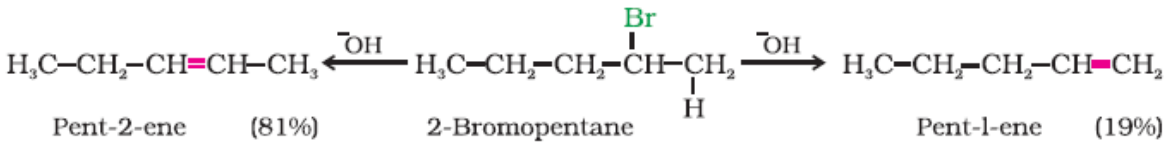


2. Elimination reactions

When a haloalkane with β-hydrogen atom is heated with alcoholic solution of potassium hydroxide, there is elimination of hydrogen atom from β-carbon and a halogen atom from the α-carbon atom. As a result, an alkene is formed as a product. Since β-hydrogen atom is involved in elimination, it is often called β-**elimination**.



If there is possibility of formation of more than one alkene due to the availability of more than one α-hydrogen atoms, usually one alkene is formed as the major product. These form part of a pattern first observed by Russian chemist, Alexander Zaitsev (also pronounced as Saytzeff) who in 1875 formulated a rule which can be summarised as "*in dehydrohalogenation reactions, the preferred product is that alkene which has the greater number of alkyl groups attached to the doubly bonded carbon atoms.*" Thus, 2-bromopentane gives pent-2-ene as the major product.

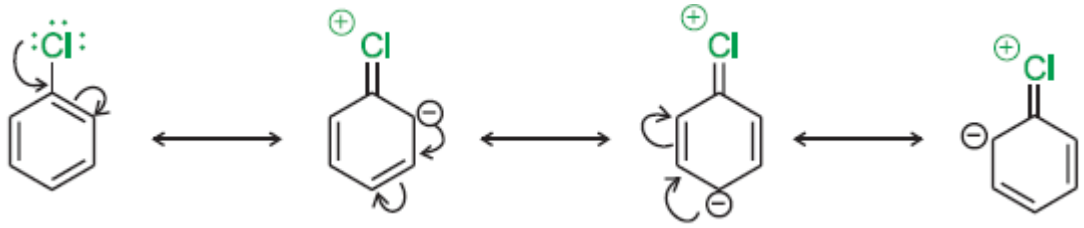


11.6.2 Reactions of Haloarenes

1. Nucleophilic substitution

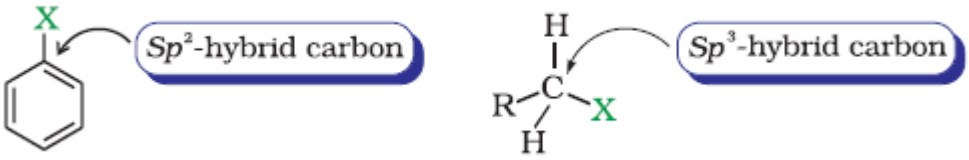
Aryl halides are extremely less reactive towards nucleophilic substitution reactions due to the following reasons:

(i) *Resonance effect :* In haloarenes, the electron pairs on halogen atom are in conjugation with □-electrons of the ring and thefollowing resonating structures are possible.



C—Cl bond acquires a partial double bond character due to resonance. As a result, the bond cleavage in haloarene is difficult than haloalkane and therefore, they are less reactive towards nucleophilic substitution reaction.

(ii) *Difference in hybridisation of carbon atom in C—X bond:* In haloalkane, the carbon atom attached to halogen is *sp3* hybridised while in case of haloarene, the carbon atom attached to halogen is *sp2*-hybridised.



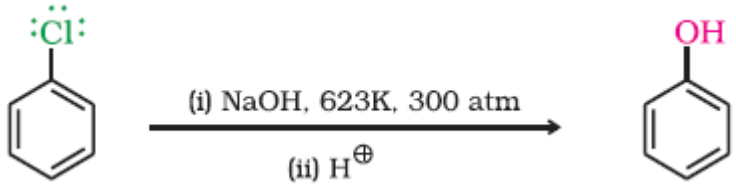
The *sp2* hybridised carbon with a greater *s*-character is more electronegative and can hold the electron pair of C—X bond more tightly than *sp3*-hybridised carbon in haloalkane with less *s*-chararcter. Thus, C—Cl bond length in haloalkane is 177pm while in haloarene is 169 pm. Since it is difficult to break a shorter bond than a longer bond, therefore, haloarenes are less reactive than haloalkanes towards nucleophilic substitution reaction.

(iii) *Instability of phenyl cation:* In case of haloarenes, the phenyl cation formed as a result of self- ionisation will not be stabilised by resonance and therefore, SN1 mechanism is ruled out.

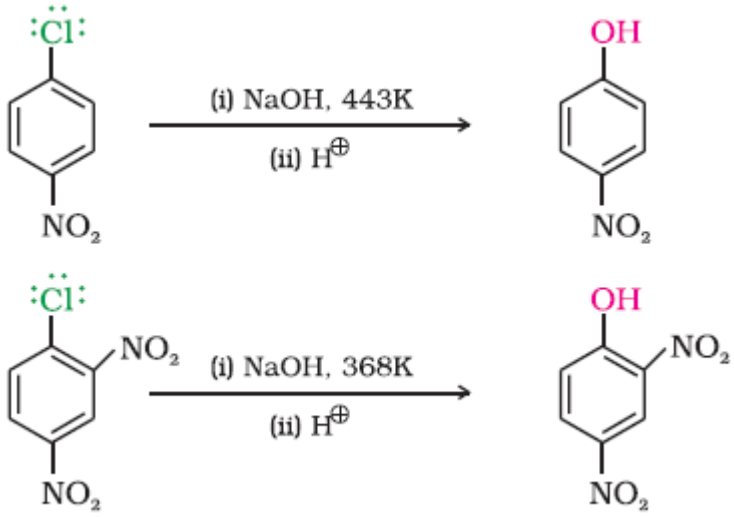
(iv) Because of the possible repulsion, it is less likely for the electron rich nucleophile to approach electron rich arenes.

Replacement by hydroxyl group

Chlorobenzene can be converted into phenol by heating in aqueous sodium hydroxide solution at a temperature of 623K and a pressure of 300 atmospheres.



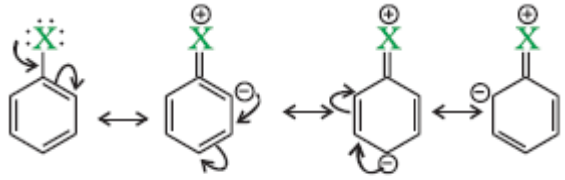
The presence of an electron withdrawing group (-NO2) at *ortho*- and *para*-positions increases the reactivity of haloarenes.



2. Electrophilic substitution reactions

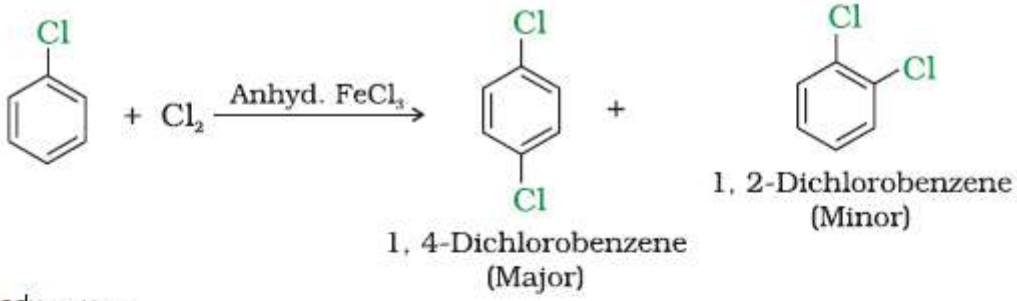
Haloarenes undergo the usual electrophilic reactions of the benzene ring such as halogenation, nitration, sulphonation and Friedel-Crafts reactions. Halogen atom besides being slightly deactivating is *o, p*directing; therefore, further substitution occurs at *ortho*- and

*para*positions with respect to the halogen atom. The *o, p*-directing influence of halogen atom can be easily understood if we consider the resonating structures of halobenzene as shown:



I II III IV

Due to resonance, the electron density increases more at *ortho*- and *para*-positions than at *meta*-positions. Further, the halogen atom because of its –I effect has some tendency to withdraw electrons from the benzene ring. As a result, the ring gets somewhat deactivated as compared to benzene and hence the electrophilic substitution reactions in haloarenes occur slowly and require more drastic conditions as compared to those in benzene.



(ii) Nitration

(iii) Sulphonation

Summary

Alkyl/ Aryl halides may be classified as mono, di, or polyhalogen (tri-, tetra-, etc.) compounds depending on whether they contain one, two or more halogen atoms in their structures. Since halogen atoms are more electronegative than carbon, the carbonhalogen bond of alkyl halide is polarised; the carbon atom bears a partial positive charge, and the halogen atom bears a partial negative charge.

Alkyl halides are prepared by the **free radical halogenation** of alkanes, addition of halogen acids to alkenes, replacement of –OH group of alcohols with halogens using

phosphorus halides, thionyl chloride or halogen acids. Aryl halides are prepared by **electrophilic substitution** to arenes. Fluorides and iodides are best prepared by halogen exchange method. The boiling points of organohalogen compounds are comparatively higher than the corresponding hydrocarbons because of strong dipole-dipole and van der Waals forces of attraction. These are slightly soluble in water but completely soluble in organic solvents.

The polarity of carbon-halogen bond of alkyl halides is responsible for their **nucleophilic substitution, elimination** and their reaction with metal atoms to form **organometallic compounds**. Nucleophilic substitution reactions are categorised into **SN1** and **SN2** on the basis of their kinetic properties. **Chirality** has a profound role in understanding the reaction mechanisms of SN1 and SN2 reactions. SN2 reactions of chiral alkyl halides are characterised by the inversion of configuration while SN1 reactions are characterised by racemisation. A number of polyhalogen compounds e.g., **dichloromethane**, **chloroform**, **iodoform**, **carbon tetrachloride**, **freon** and **DDT** have many industrial applications. However, some of these compounds cannot be easily decomposed and even cause depletion of ozone layer and are proving **environmental hazards**.

CHAPTER 12

Alcohols, Phenols and Ethers

An alcohol contains one or more hydroxyl (OH) group(s) directly attached to carbon atom(s), of an aliphatic system (CH_3OH) while a phenol contains –OH group(s) directly attached to carbon atom(s) of an aromatic system (C_6H_5OH).

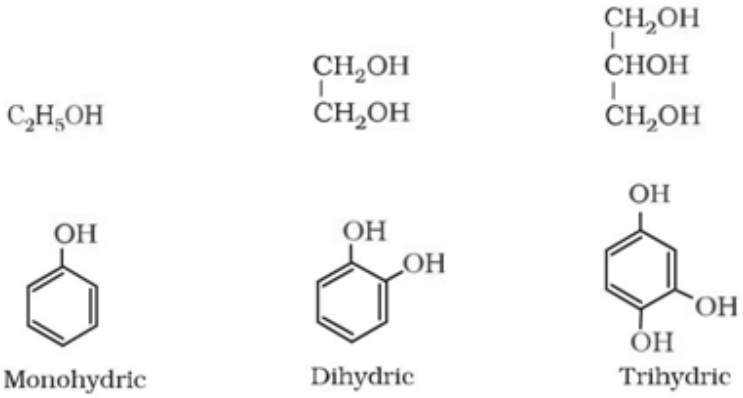
The substitution of a hydrogen atom in a hydrocarbon by an alkoxy or aryloxy group (R–O/Ar–O) yields another class of compounds known as 'ethers', for example, CH_3OCH_3 (dimethyl ether). You may also visualise ethers as compounds formed by substituting the hydrogen atom of hydroxyl group of an alcohol or phenol by an alkyl or aryl group.

12.1 Classification

The classification of compounds makes their study systematic and hence simpler. Therefore, let us first learn how are alcohols, phenols and ethers classified?

Mono, Di, Tri or Polyhydric Compounds

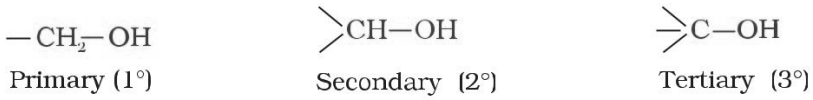
Alcohols and phenols may be classified as mono–, di–, tri- or polyhydric compounds depending on whether they contain one, two, three or many hydroxyl groups respectively in their structures as given below:



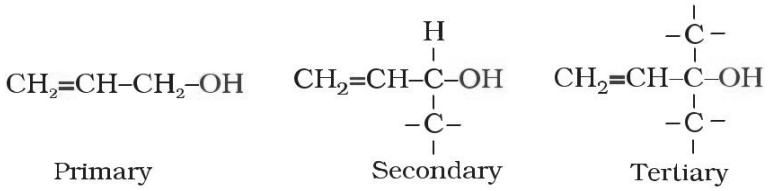
Monohydric alcohols may be further classified according to the hybridisation of the carbon atom to which the hydroxyl group is attached.

(i) Compounds containing $C_{sp}3 - OH$ *bond:* In this class of alcohols, the –OH group is attached to an sp^3 hybridised carbon atom of an alkyl group. They are further classified as follows:

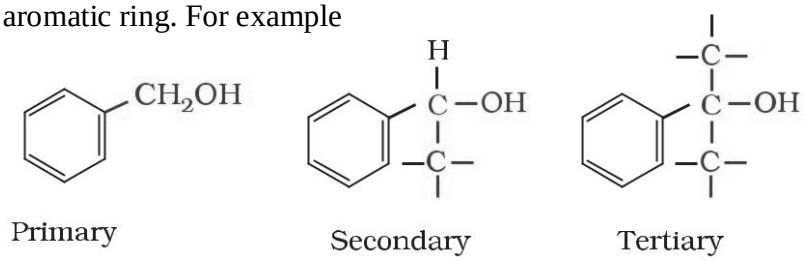
Primary, secondary and tertiary alcohols: In these three types of alcohols, the –OH group is attached to primary, secondary and tertiary carbon atom, respectively as depicted below:



Allylic alcohols: In these alcohols, the —OH group is attached to a sp^3 hybridised carbon next to the carbon-carbon double bond, that is to an allylic carbon. For example

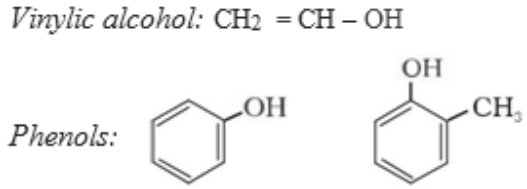


Benzylic alcohols*:* In these alcohols, the —OH group is attached to a sp^3—hybridised carbon atom next to an aromatic ring. For example



Allylic and benzylic alcohols may be primary, secondary or tertiary.

(ii) Compounds containing $C_{sp}2$ − OH bond: These alcohols contain —OH group bonded to a carbon-carbon double bond i.e., to a vinylic carbon or to an aryl carbon. These alcohols are also known as vinylic alcohols.



Ethers are classified as **simple** or **symmetrical**, if the alkyl or aryl groups attached to the oxygen atom are the same, and **mixed** or **unsymmetrical**, if the two groups are different. Diethyl ether, $C_2H_5OC_2H_5$, is a symmetrical ether whereas $C_2H_5OCH_3$ and $C_2H_5OC_6H_5$ are unsymmetrical ethers.

12.2 Nomenclature (Alcohols, Ethers & Phenols)

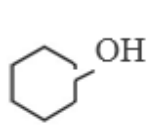
(a) Alcohols: The common name of an alcohol is derived from the common name of the alkyl group and adding the word alcohol to it. For example, CH_3OH is methyl alcohol.

According to IUPAC system (Unit 12, Class XI), the name of an alcohol is derived from the name of the alkane from which the alcohol is derived, by substituting 'e' of alkane with the suffix 'ol'. The position of substituents is indicated by numerals. For this, the longest carbon chain (parent chain) is numbered starting at the end nearest to the hydroxyl group. The positions of the –OH group and other substituents are indicated by using the numbers of carbon atoms to which these are attached. For naming polyhydric alcohols, the 'e' of alkane is retained and the ending 'ol' is added. The number of –OH groups is indicated by adding the multiplicative prefix, di, tri, etc., before 'ol'. The positions of –OH groups are indicated by appropriate locants e.g., $HO–CH_2–CH_2–OH$ is named as ethane–1, 2-diol. The following Table gives common and IUPAC names of a few alcohols as examples.

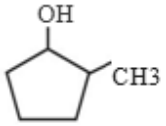
Common and IUPAC names of some Alcohols

Compound	Common name	Common name
CH_3-OH	Methyl alcohol	Methanol
$CH_3-CH_2-CH_2-OH$	*n*-Propyl alcohol	Propan-1-ol
$CH_3-CH(OH)-CH_3$	Isopropyl alcohol	Propan-2-ol
$CH_3-CH_2-CH_2-CH_2-OH$	*n*-Butyl alcohol	Butan-1-ol
$CH_3-CH(OH)-CH_2-CH_3$	*sec*-Butyl alcohol	Butan-2-ol
$CH_3-CH(OH)-CH_2-CH_3$	Isobutyl alcohol	2-Methylpropan-1-ol
$CH_3-C(CH_3)(OH)-CH_3$	*tert*-Butyl alcohol	2-Methylpropan-2-ol
$CH_2(OH)-CH(OH)-CH_2(OH)$	Glycerol	Propane -1, 2, 3-triol

Cyclic alcohols are named using the prefix cyclo and considering the –OH group attached to C–1.



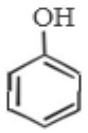
Cyclohexanol



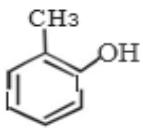
2-Methylcyclopentanol

(b) Phenols: The simplest hydroxy derivative of benzene is phenol. It is its common name and also an accepted IUPAC name. As structure of phenol involves a benzene ring, in its substituted compounds the terms *ortho* (1,2- disubstituted), *meta* (1,3-disubstituted) and *para* (1,4-disubstituted) are often used in the common names.

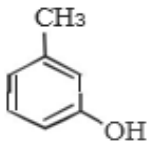
Common name IUPAC name



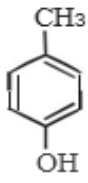
Phenol
Phenol



o-Cresol
2-Methylphenol



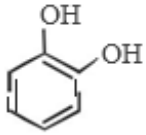
m-Cresol
3-Methylphenol



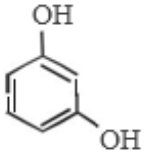
p-Cresol
4-Methylphenol

Dihydroxy derivatives of benzene are known as 1, 2-, 1, 3- and 1, 4-benzenediol.

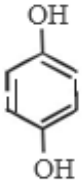
Common name IUPAC name



Catechol
Benzene-1,2-diol



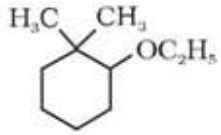
Resorcinol
Benzene-1,3-diol



Hydroquinone or quinol
Benzene-1,4-diol

(c) Ethers: Common names of ethers are derived from the names of alkyl/ aryl groups written as separate words in alphabetical order and adding the word 'ether' at the end. For example, $CH_3OC_2H_5$ is ethylmethyl ether.

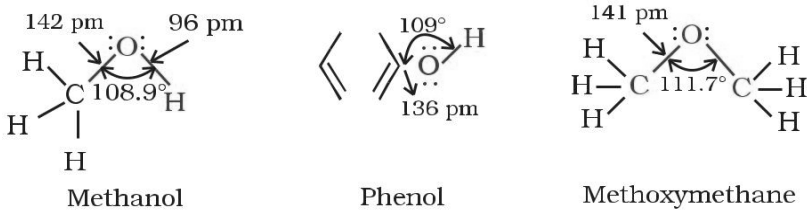
Common and IUPAC names of some Ethers

Compound	Common name	IUPAC name
CH_3OCH_3	Dimethyl ether	Methoxymethane
$C_2H_5OC_2H_5$	Diethyl ether	Ethoxyethane
$CH_3OCH_2CH_2CH_3$	Methyl n-propyl ether	1-Methoxypropane
$C_6H_5OCH_3$	Methylphenyl ether (Anisole)	Methoxybenzene (Anisole)
$C_6H_5OCH_2CH_3$	Ethylphenyl ether (Phenetole)	Ethoxybenzene
$C_6H_5O(CH_2)_6-CH_3$	Heptylphenyl ether	1-Phenoxyheptane
$CH_3O-\underset{\displaystyle CH_3}{\overset{\displaystyle CH-CH_3}{|}}$	Methyl isopropyl ether	2-Methoxypropane
$C_6H_5-O-CH_2-CH_2-\underset{\displaystyle CH_3}{\overset{\displaystyle CH-CH_3}{|}}$	Phenylisopentyl ether	3- Methylbutoxybenzene
$CH_3-O-CH_2-CH_2-OCH_3$	—	1,2-Dimethoxyethane
	—	2-Ethoxy- -1,1-dimethylcyclohexane

If both the alkyl groups are the same, the prefix 'di' is added before the alkyl group. For example, $C_2H_5OC_2H_5$ is diethyl ether. According to IUPAC system of nomenclature, ethers are regarded as hydrocarbon derivatives in which a hydrogen atom is replaced by an–OR or –OAr group, where R and Ar represent alkyl and aryl groups, respectively. The larger (R) group is chosen as the parent hydrocarbon. The names of a few ethers are given as examples in the above Table.

12.3 Structures of Groups

In alcohols, the oxygen of the –OH group is attached to carbon by a sigma (σ) bond formed by the overlap of a sp^3 hybridised orbital of structural aspects of methanol, phenol and methoxymethane.



Structures of methanol, phenol and methoxymethane

The bond angle in alcohols is slightly less than the tetrahedral angle (109°-28′). It is due to the repulsion between the unshared electron pairs of oxygen. In phenols, the –OH group is attached to sp^2 hybridised carbon of an aromatic ring. The carbon– oxygen bond length (136 pm) in phenol is slightly less than that in methanol. This is due to (i) partial double bond character on account of the conjugation of unshared electron pair of oxygen with the aromatic ring (i) and (ii) sp^2 hybridised state of carbon to which oxygen is attached.

In ethers, the four electron pairs, i.e., the two bond pairs and two lone pairs of electrons on oxygen are arranged approximately in a tetrahedral arrangement. The bond angle is slightly greater than the tetrahedral angle due to the repulsive interaction between the two bulky (–R) groups. The C–O bond length (141 pm) is almost the same as in alcohols.

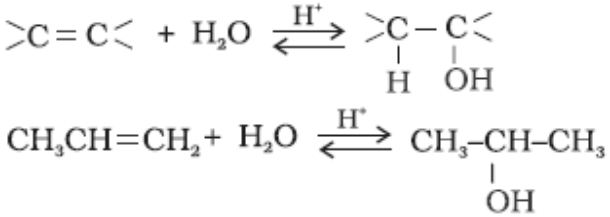
12.4 Preparation of Alcohols and Phenols

Preparation of Alcohols

Alcohols are prepared by the following methods:

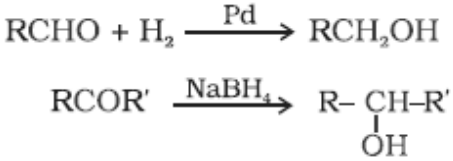
1. From alkenes

By acid catalysed hydration*:* Alkenes react with water in the presence of acid as catalyst to form alcohols. In case of unsymmetrical alkenes, the addition reaction takes place in accordancewith Markovnikov's rule



2. From carbonyl compounds

By reduction of aldehydes and ketones: Aldehydes and ketones are reduced to the corresponding alcohols by addition of hydrogen in the presence of catalysts (catalytic hydrogenation). The usual catalyst is a finely divided metal such as platinum, palladium or nickel. It is also prepared by treating aldehydes and ketones with sodium borohydride (NaBH4) or lithium aluminium hydride (LiAlH4). Aldehydes yield primary alcohols whereas ketones give secondary alcohols.

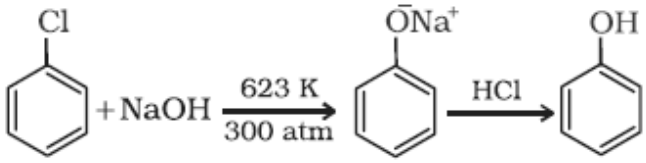


Preparation of Phenols

Phenol, also known as carbolic acid, was first isolated in the early nineteenth century from coal tar. Nowadays, phenol is commercially produced synthetically. In the laboratory, phenols are prepared from benzene derivatives by any of the following methods:

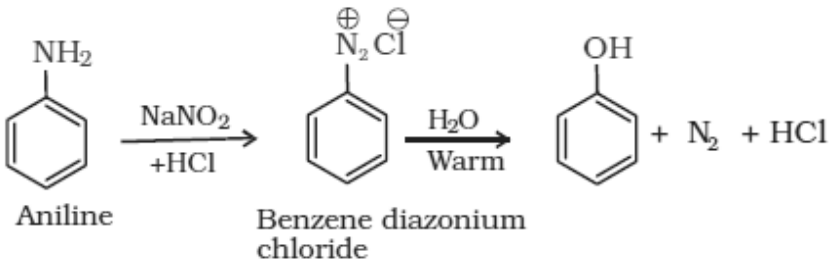
From haloarenes

Chlorobenzene is fused with NaOH at 623K and 320 atmospheric pressure. Phenol is obtained by acidification of sodium phenoxide so produced



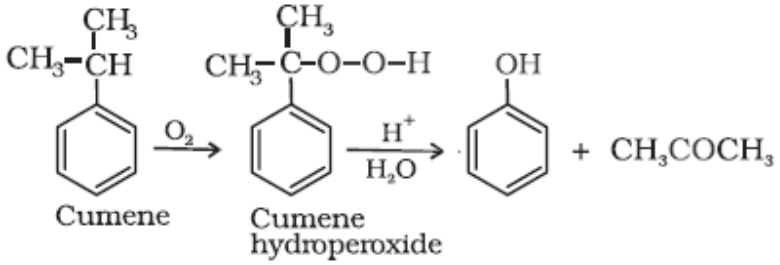
From diazonium salts

A diazonium salt is formed by treating an aromatic primary amine with nitrous acid (NaNO2 + HCl) at 273-278 K. Diazonium salts are hydrolysed to phenols by warming with water or by treating with dilute acids



From cumene

Phenol is manufactured from the hydrocarbon, cumene. Cumene (isopropylbenzene) is oxidised in the presence of air to cumene hydroperoxide. It is converted to phenol and acetone by treating it with dilute acid. Acetone, a by-product of this reaction, is also obtained in large quantities by this method.



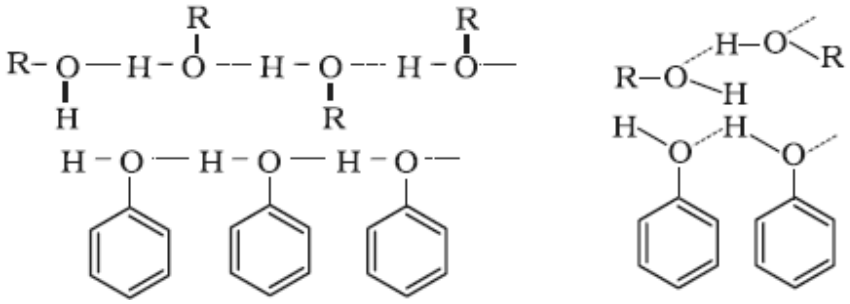
12.5 Physical Properties

Alcohols and phenols consist of two parts, an alkyl/aryl group and a hydroxyl group. The properties of alcohols and phenols are chiefly due to the hydroxyl group. The nature of alkyl and aryl groups simply modify these properties.

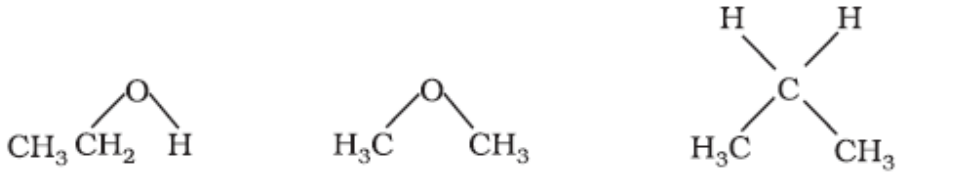
Boiling Points

The boiling points of alcohols and phenols increase with increase in the number of carbon atoms (increase in van der Waals forces). In alcohols, the boiling points decrease with increase of branching in carbon chain (because of decrease in van der Waals forces with decrease in surface area).

The –OH group in alcohols and phenols is involved in intermolecular hydrogen bonding as shown below:



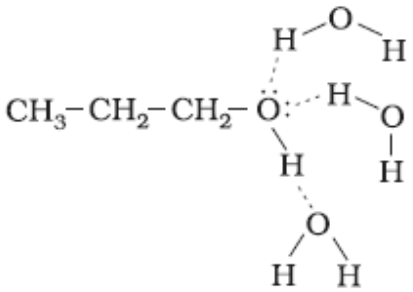
It is interesting to note that boiling points of alcohols and phenols are higher in comparison to other classes of compounds, namely hydrocarbons, ethers, haloalkanes and haloarenes of comparable molecular masses. For example, ethanol and propane have comparable molecular masses but their boiling points differ widely. The boiling point of methoxymethane is intermediate of the two boiling points.



The high boiling points of alcohols are mainly due to the presence of intermolecular hydrogen bonding in them which is lacking in ethers and hydrocarbons.

Solubility

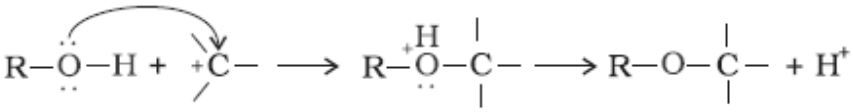
Solubility of alcohols and phenols in water is due to their ability to form hydrogen bonds with water molecules as shown. The solubility decreases with increase in size of alkyl/aryl (hydrophobic) groups. Several of the lower molecular mass alcohols are miscible with water in all proportions.



12.6 Chemical Reactions of Alcohols and Phenols

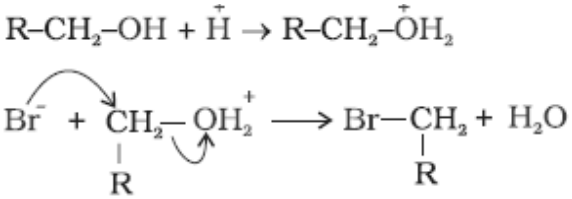
Alcohols are versatile compounds. They react both as nucleophiles and electrophiles. The bond between O–H is broken when alcohols react as nucleophiles.

(a) Alcohols as nucleophiles



(b) **Protonated alcohols as electrophiles**

The bond between C–O is broken when they react as electrophiles. Protonated alcohols react in this manner.

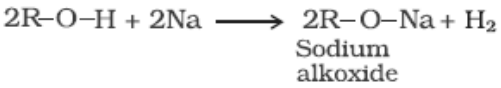


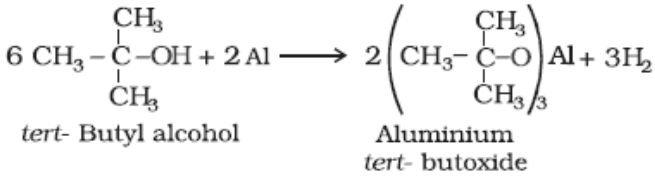
Based on the cleavage of O–H and C–O bonds, the reactions of alcohols and phenols may be divided into two groups:

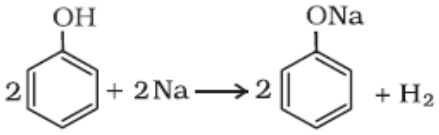
12.7 Reactions involving Cleavage of -OH bond:

(a) **Acidity of alcohols and phenols**

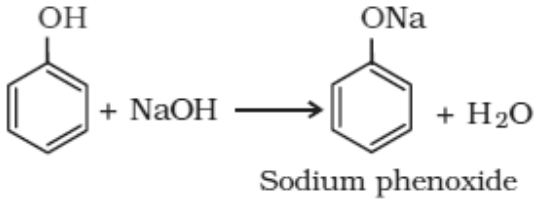
Reaction with metals: Alcohols and phenols react with active metals such as sodium, potassium and aluminium to yield corresponding alkoxides/phenoxides and hydrogen.



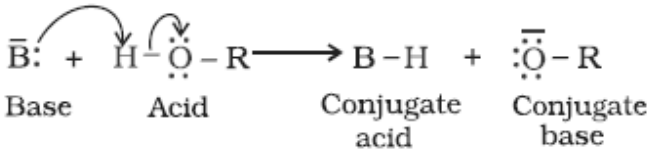




In addition to this, phenols react with aqueous sodium hydroxide to form sodium phenoxides.

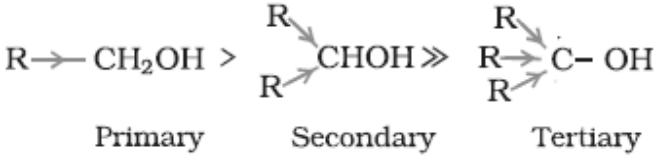


The above reactions show that alcohols and phenols are acidic in nature. In fact, alcohols and phenols are Brönsted acids i.e., they can donate a proton to a stronger base (B:).

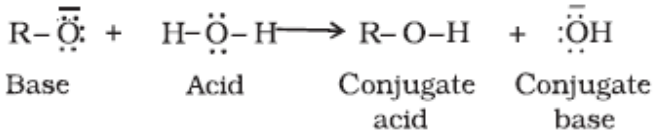


(b) Acidity of alcohols:

The acidic character of alcohols is due to the polar nature of O–H bond. An electron-releasing group ($–CH_3$, $–C_2H_5$) increases electron density on oxygen tending to decrease the polarity of O-H bond. This decreases the acid strength. For this reason, the acid strength of alcohols decreases in the following order:



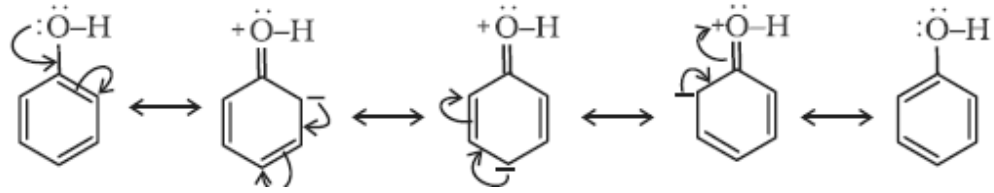
Alcohols are, however, weaker acids than water. This can be illustrated by the reaction of water with an alkoxide.



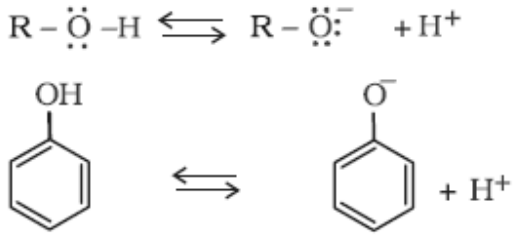
This reaction shows that water is a better proton donor (i.e., stronger acid) than alcohol.

(c) Acidity of phenols:

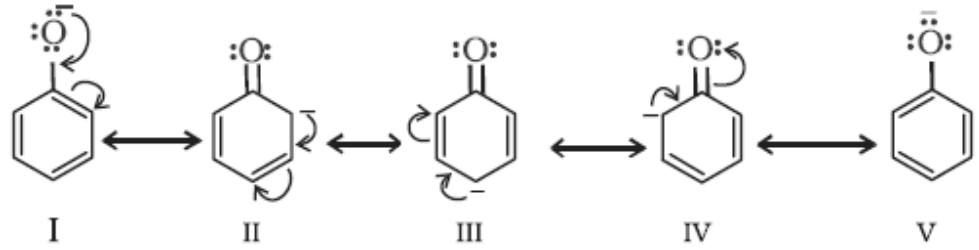
The reactions of phenol with metals (e.g., sodium, aluminum) and sodium hydroxide indicate its acidic nature. The hydroxyl group, in phenol is directly attached to the sp^2 hybridised carbon of benzene ring which acts as an electron withdrawing group. Due to this, the charge distribution in phenol molecule, as depicted in its resonance structures, causes the oxygen of –OH group to be positive.



The reaction of phenol with aqueous sodium hydroxide indicates that phenols are stronger acids than alcohols and water. Let us examine how a compound in which hydroxyl group attached to an aromatic ring is more acidic than the one in which hydroxyl group is attached to an alkyl group. The ionisation of an alcohol and a phenol takes place as follows:



Due to the higher electronegativity of sp^2 hybridised carbon of phenol to which –OH is attached, electron density decreases on oxygen. This increases the polarity of O–H bond and results in an increase in ionisation of phenols than that of alcohols. Now let us examine the stabilities of alkoxide and phenoxide ions. In alkoxide ion, the negative charge is localised on oxygen while in phenoxide ion, the charge is delocalised. The delocalisation of negative charge (structures I-V) makes phenoxide ion more stable and favours the ionisation of phenol. Although there is also charge delocalisation in phenol, its resonance structures have charge separation due to which the phenol molecule is less stable than phenoxide ion.



In substituted phenols, the presence of electron withdrawing groups such as nitro group, enhances the acidic strength of phenol. This effect is more pronounced when such a group is present at *ortho* and *para* positions. It is due to the effective delocalisation of negative charge in phenoxide ion. On the other hand, electron releasing groups, such as alkyl groups, in general, do not favour the formation of phenoxide ion resulting in decrease in acid strength. Cresols, for example, are less acidic than phenol.

Table p*Ka* Values of some Phenols and Ethanol

Compound	Formula	pK_a
o-Nitrophenol	*o*–O_2N–C_6H_4–OH	7.2
m-Nitrophenol	*m*–O_2N–C_6H_4–OH	8.3
p-Nitrophenol	*p*-O_2N–C_6H_4–OH	7.1
Phenol	C_6H_5–OH	10.0
o-Cresol	*o*-CH_3–C_6H_4–OH	10.2
m-Cresol	*m*-$CH_3C_6H_4$–OH	10.1
p-Cresol	*p*-CH_3–C_6H_4–OH	10.2
Ethanol	C_2H_5OH	15.9

From the above data, you will note that phenol is million times more acidic than ethanol.

e. Esterification

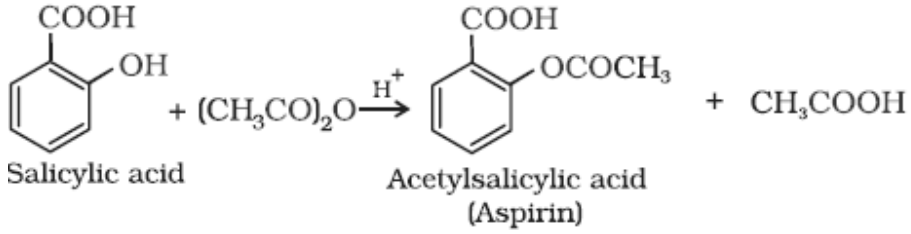
Alcohols and phenols react with carboxylic acids, acid chlorides and acid anhydrides to form esters.

$$Ar/RO-H + R'-COOH \overset{H^+}{\rightleftharpoons} Ar/ROCOR' + H_2O$$

$$Ar/R-OH + (R'CO)_2O \overset{H^+}{\rightleftharpoons} Ar/ROCOR' + R'COOH$$

$$R/ArOH + R'COCl \xrightarrow{\text{Pyridine}} R/ArOCOR' + HCl$$

The reaction with carboxylic acid and acid anhydride is carried out in the presence of a small amount of concentrated sulphuric acid. The reaction is reversible, and therefore, water is removed as soon as it is formed. The reaction with acid chloride is carried out in the presence of a base (pyridine) so as to neutralise HCl which is formed during the reaction. It shifts the equilibrium to the right hand side. The introduction of acetyl (CH_3CO) group in alcohols or phenols is known as acetylation. Acetylation of salicylic acid produces aspirin.



12.8 Reactions involving cleavage of carbon – oxygen (C–O) bond in alcohols

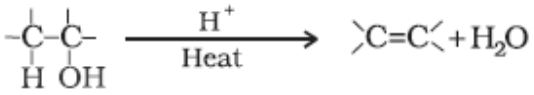
The reactions involving cleavage of C–O bond take place only in alcohols. Phenols show this type of reaction only with zinc. ***1. Reaction with hydrogen halides:*** Alcohols react with hydrogen halides to form alkyl halides

$$ROH + HX \rightarrow R\text{–}X + H_2O$$

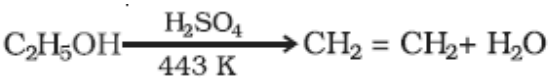
The difference in reactivity of three classes of alcohols with HCl distinguishes them from one another (**Lucas test**). Alcohols are soluble in Lucas reagent (conc. HCl and $ZnCl_2$) while their halides are immiscible and produce turbidity in solution. In case of tertiary alcohols, turbidity is produced immediately as they form the halides easily. Primary alcohols do not produce turbidity at room temperature.

a) Dehydration:

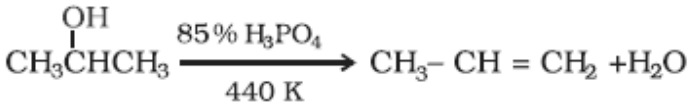
Alcohols undergo dehydration (removal of a molecule of water) to form alkenes on treating with a protic acid e.g., concentrated H_2SO_4 or H_3PO_4, or catalysts such as anhydrous zinc chloride or alumina.

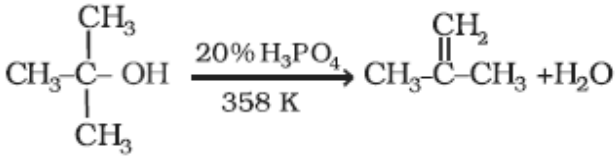


Ethanol undergoes dehydration by heating it with concentrated H_2SO_4 at 443 K.



Secondary and tertiary alcohols are dehydrated under milder conditions. For example



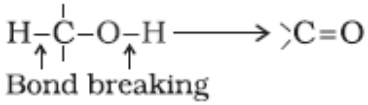


Thus, the relative ease of dehydration of alcohols follows the following order:

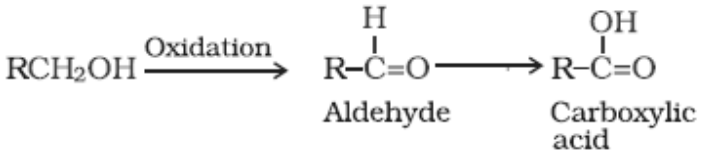
Tertiary > Secondary > Primary

b) Oxidation:

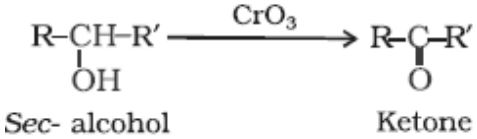
Oxidation of alcohols involves the formation of a carbonoxygen double bond with cleavage of an O-H and C-H bonds.



Such a cleavage and formation of bonds occur in oxidation reactions. These are also known as **dehydrogenation** reactions as these involve loss of dihydrogen from an alcohol molecule. Depending on the oxidising agent used, a primary alcohol is oxidised to an aldehyde which in turn is oxidised to a carboxylic acid.



Secondary alcohols are oxidised to ketones by chromic anhydride (CrO_3).



Tertiary alcohols do not undergo oxidation reaction. Under strong reaction conditions such as strong oxidising agents ($KMnO_4$) and elevated temperatures, cleavage of various C-C bonds takes place and a mixture of carboxylic acids containing lesser number of carbon atoms is formed.

12.9 Reactions of phenols

Following reactions are shown by phenols only.

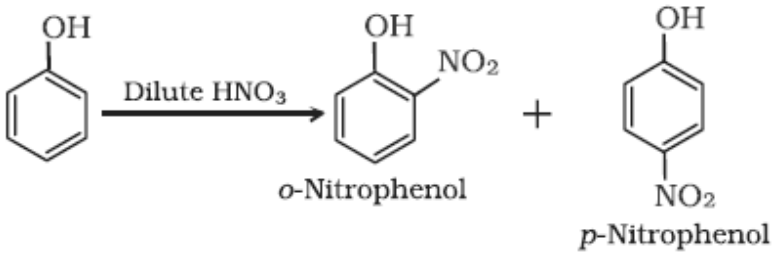
a) Electrophilic aromatic substitution

In phenols, the reactions that take place on the aromatic ring are electrophilic substitution reactions (Unit 13, Class XI). The –OH group attached to the benzene ring activates it towards electrophilic substitution. Also, it directs the incoming group to *ortho* and *para* positions in the ring as these positions become electron rich due to the resonance effect caused by –OH group. The resonance structures are shown under acidity of phenols.

Common electrophilic aromatic substitution reactions taking place in phenol are as follows:

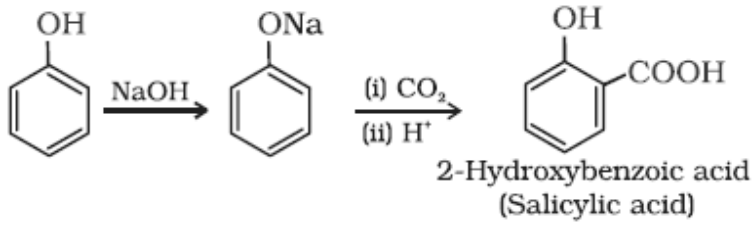
1. Nitration:

With dilute nitric acid at low temperature (298 K), phenol yields a mixture of *ortho* and *para* nitrophenols.



2. Kolbe's reaction

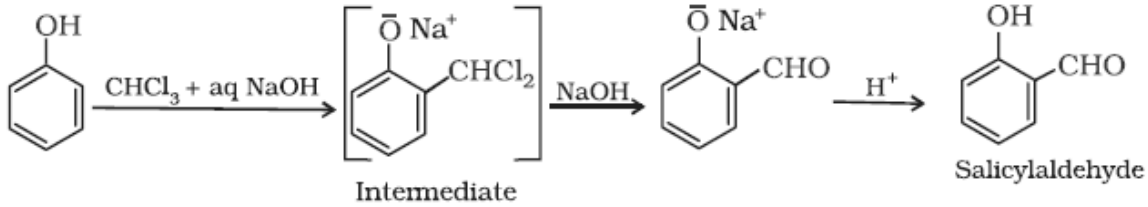
Phenoxide ion generated by treating phenol with sodium hydroxide is even more reactive than phenol towards electrophilic aromatic substitution. Hence, it undergoes electrophilic substitution with carbon dioxide, a weak electrophile. *Ortho* hydroxybenzoic acid is formed as the main reaction product.



2-Hydroxybenzoic acid
(Salicylic acid)

3. Reimer-Tiemann reaction

On treating phenol with chloroform in the presence of sodium hydroxide, a –CHO group is introduced at *ortho* position of benzene ring. This reaction is known as *Reimer - Tiemann reaction*. The intermediate substituted benzal chloride is hydrolysed in the presence of alkali to produce salicylaldehyde.

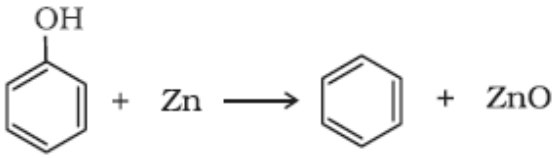


Intermediate

Salicylaldehyde

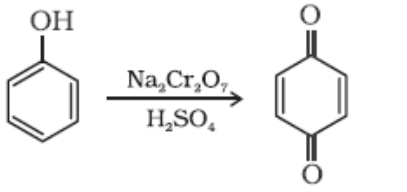
4. Reaction of phenol with zinc dust

Phenol is converted to benzene on heating with zinc dust.



5. Oxidation

Oxidation of phenol with chromic acid produces a conjugated diketone known as benzoquinone. In the presence of air, phenols are slowly oxidised to dark coloured mixtures containing quinones.



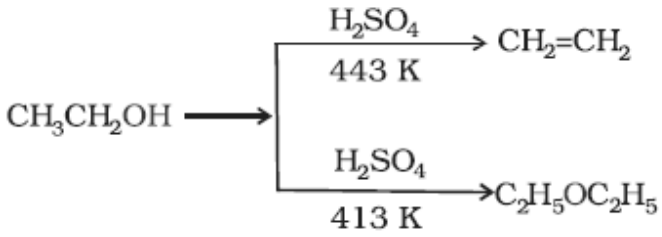
benzoquinone

12.10 Ethers

Preparation of Ethers

1. **By dehydration of alcohols**

Alcohols undergo dehydration in the presence of protic acids (H_2SO_4, H_3PO_4). The formation of the reaction product, alkene or ether depends on the reaction conditions. For example, ethanol is dehydrated to ethene in the presence of sulphuric acid at 443 K. At 413 K, ethoxyethane is the main product.



2. Williamson synthesis

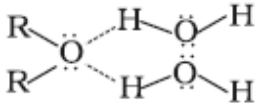
It is an important laboratory method for the preparation of symmetrical and unsymmetrical ethers. In this method, an alkyl halide is allowed to react with sodium alkoxide.

Physical Properties

The C-O bonds in ethers are polar and thus, ethers have a net dipole moment. The weak polarity of ethers do not appreciably affect their boiling points which are comparable to those of the alkanes of comparable molecular masses but are much lower than the boiling points of alcohols as shown in the following cases:

Formula	$CH_3(CH_2)_3CH_3$ n-Pentane	C_2H_5-O-C_2H_5 Ethoxyethane	$CH_3(CH_2)_3$-OH Butan-1-ol
b.p./K	309.1	307.6	390

The large difference in boiling points of alcohols and ethers is due to the presence of hydrogen bonding in alcohols. The miscibility of ethers with water resembles those of alcohols of the same molecular mass. Both ethoxyethane and butan-1-ol are miscible to almost the same extent i.e., 7.5 and 9 g per 100 mL water, respectively while pentane is essentially immiscible with water. Can you explain this observation ? This is due to the fact that just like alcohols, oxygen of ether can also form hydrogen bonds with water molecule as shown:



Chemical Reactions

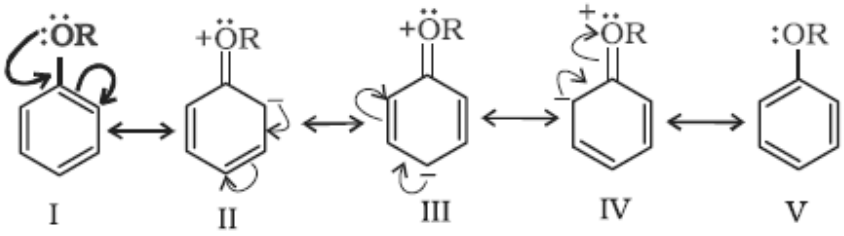
1. Cleavage of C–O bond in ethers

Ethers are the least reactive of the functional groups. The cleavage of C-O bond in ethers takes place under drastic conditions with excess of hydrogen halides. The reaction of dialkyl ether gives two alkyl halide molecules.

Alkyl aryl ethers are cleaved at the alkyl-oxygen bond due to the more stable aryl-oxygen bond. The reaction yields phenol and alkyl halide.

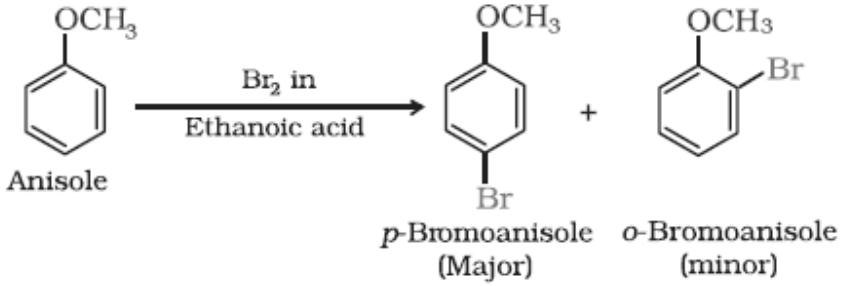
2. Electrophilic substitution

The alkoxy group (-OR) is *ortho*, *para* directing and activates the aromatic ring towards electrophilic substitution in the same way as in phenol.



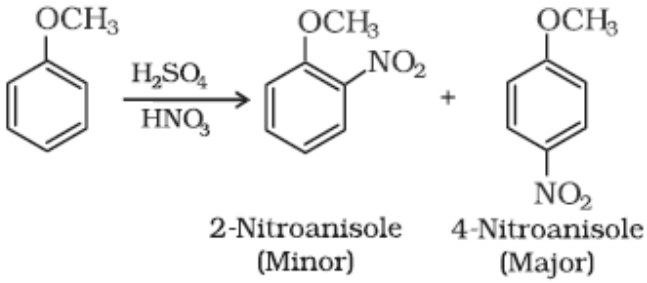
(i) Halogenation:

Phenylalkyl ethers undergo usual halogenations in the benzene ring, *e.g.*, anisole undergoes bromination with bromine in ethanoic acid even in the absence of iron (III) bromide catalyst. It is due to the activation of benzene ring by the methoxy group. *Para* isomer is obtained in 90% yield.



(ii) Nitration:

Anisole reacts with a mixture of concentrated sulphuric and nitric acids to yield a mixture of *ortho* and *para* nitroanisole.



Summary

Alcohols and **phenols** are classified (i) on the basis of the number of hydroxyl groups and (ii) according to the hybridisation of the carbon atom, sp^3 or sp^2 to which the –OH group is attached. **Ethers** are classified on the basis of groups attached to the oxygen atom.

Alcohols may be prepared (1) by hydration of alkenes (i) in presence of an acid and (ii) by hydroboration-oxidation reaction (2) from carbonyl compounds by (i) catalytic reduction and (ii) the action of Grignard reagents. Phenols may be prepared by (1) substitution of (i) halogen atom in haloarenes and (ii) sulphonic acid group in aryl sulphonic acids, by –OH group (2) by hydrolysis of diazonium salts and (3) industrially from cumene.

Alcohols are higher boiling than other classes of compounds, namely hydrocarbons, ethers and haloalkanes of comparable molecular masses. The ability of alcohols, phenols and ethers to form intermolecular hydrogen bonding with water makes them soluble in it.

Alcohols and phenols are acidic in nature. **Electron withdrawing groups** in phenol increase its acidic strength and **electron releasing groups** decrease it. Alcohols undergo nucleophilic substitution with hydrogen halides to yield alkyl halides. Dehydration of alcohols gives alkenes. On oxidation, primary alcohols yield aldehydes with mild oxidising agents and carboxylic acids with strong oxidising agents while secondary alcohols yield ketones. Tertiary alcohols are resistant to oxidation.

The presence of –OH group in phenols activates the aromatic ring towards **electrophilic substitution** and directs the incoming group to *ortho* and *para* positions due to resonance effect. **Reimer-Tiemann reaction** of phenol yields salicylaldehyde. In presence of sodium hydroxide, phenol generates phenoxide ion which is even more reactive than phenol. Thus, in alkaline medium, phenol undergoes **Kolbe's reaction**.

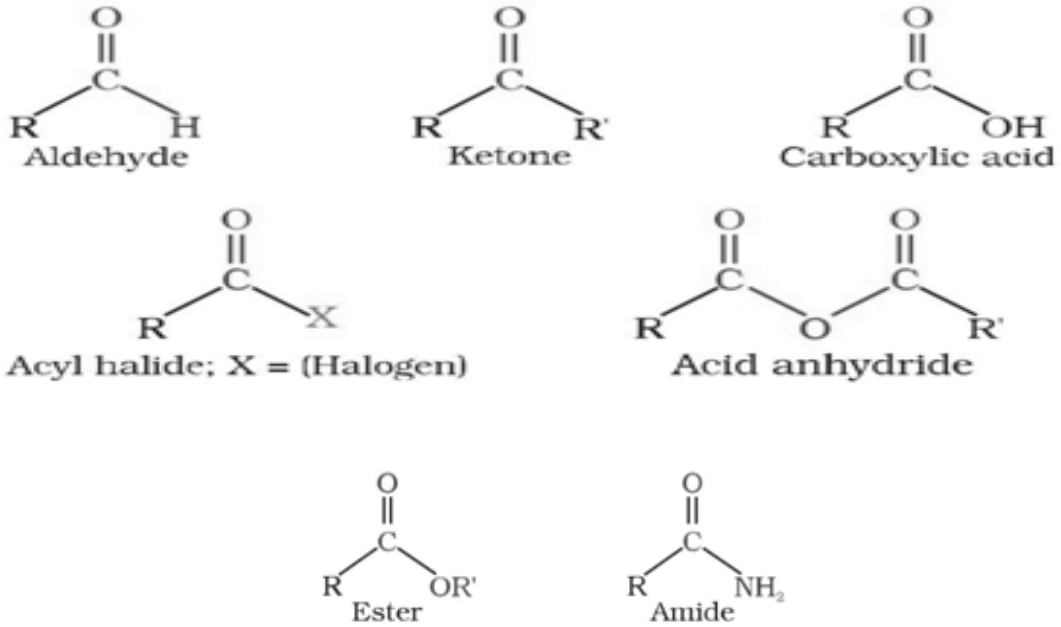
Ethers may be prepared by (i) dehydration of alcohols and (ii) **Williamson synthesis**. The boiling points of ethers resemble those of alkanes while their solubility is comparable to those of alcohols having same molecular mass. The C–O bond in ethers can be cleaved by hydrogen halides. In electrophilic substitution, the alkoxy group activates the aromatic ring and directs the incoming group to ortho and para positions.

IMPORTANT QUESTIONS

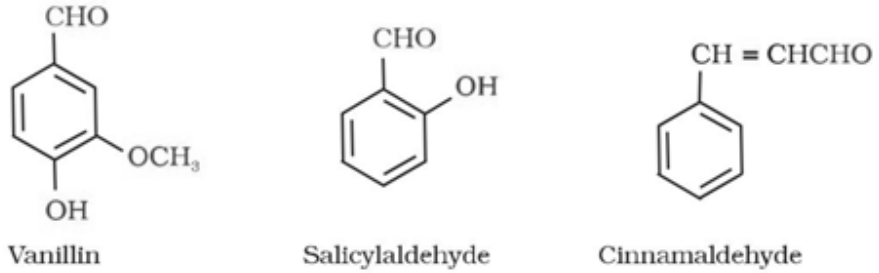
1. Write any two methods of preparations of Alcohols, Phenols and Ethers
2. Explain the Acidic nature of Phenols and compare with that of Alcohols
3. Explain the following named Equations.
    a. Kolbe's reaction
    b. Reimer – Tiemann reaction
    c. Williamsons ether synthesis.

ALDEHYDES, KETONES AND CARBOXYLIC ACIDS

In aldehydes, the carbonyl group is bonded to a carbon and hydrogen while in the ketones, it is bonded to two carbon atoms. The carbonyl compounds in which carbonyl group is bonded to oxygen are known as carboxylic acids, and their derivatives (e.g. esters, anhydrides) while in compounds where carbon is attached to nitrogen and to halogens are called amides and acyl halides respectively. The general formulas of these classes of compounds are given below:



Aldehydes, ketones and carboxylic acids are widespread in plants and animal kingdom. They play an important role in biochemical processes of life. They add fragrance and flavour to nature, for example, vanillin (from vanilla beans), salicylaldehyde (from meadow sweet) and cinnamaldehyde (from cinnamon) have very pleasant fragrances.



They are used in many food products and pharmaceuticals to add flavours. Some of these families are manufactured for use as solvents (i.e., acetone) and for preparing materials like adhesives, paints, resins, perfumes, plastics, fabrics, etc.

12.11 Nomenclature and Structure of Carbonyl Group

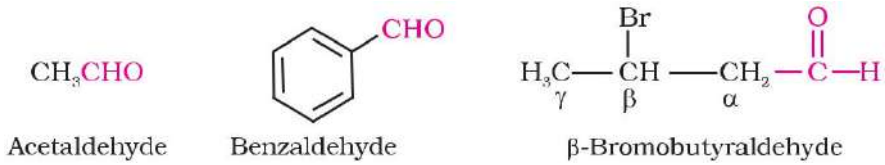
Nomenclature

I. Aldehydes and ketones

Aldehydes and ketones are the simplest and most important carbonyl compounds. There are two systems of nomenclature of aldehydes and ketones.

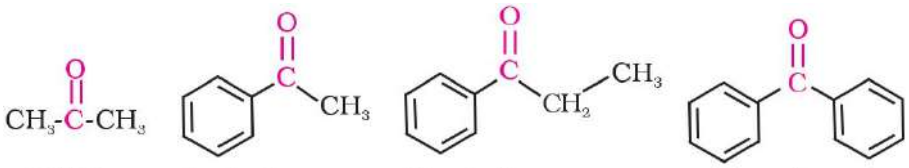
(a) Common names

Aldehydes and ketones are often called by their common names instead of IUPAC names. The common names of most aldehydes are derived from the common names of the corresponding carboxylic acids [Section 12.6.1] by replacing the ending *–ic* of acid with aldehyde. At the same time, the names reflect the Latin or Greek term for the original source of the acid or aldehyde. The location of the substituent in the carbon chain is indicated by Greek letters α, β, γ, δ, etc. The α-carbon being the one directly linked to the aldehyde group, β-carbon the next, and so on. For example



Acetaldehyde Benzaldehyde β-Bromobutyraldehyde

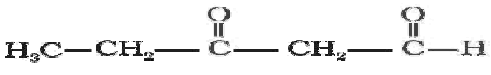
The common names of ketones are derived by naming two alkyl or aryl groups bonded to the carbonyl group. The locations of substituents are indicated by Greek letters, α α′, β β′ and so on beginning with the carbon atoms next to the carbonyl group, indicated as αα′. Some ketones have historical common names, the simplest dimethyl ketone is called acetone. Alkyl phenyl ketones are usually named by adding the acyl group as prefix to phenone. For example

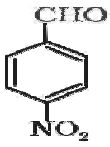


Acetone Acetophenone Propiophenone Benzophenone

(ii) IUPAC names

The IUPAC names of open chain aliphatic aldehydes and ketones are derived from the names of the corresponding alkanes by replacing the ending *–e* with *–al* and *–one* respectively. In case of aldehydes the longest carbon chain is numbered starting from the carbon of the aldehyde group while in case of ketones the numbering begins from the end nearer to the carbonyl group. The substituents are prefixed in alphabetical order along with numerals indicating their positions in the carbon chain. The same applies to cyclic ketones, where the carbonyl carbon is numbered one. When the aldehyde group is attached to a ring, the suffix carbaldehyde is added after the full name of the cycloalkane. The numbering of the ring carbon atoms start from the carbon atom attached to the aldehyde group. The name of the simplest aromatic aldehyde carrying the aldehyde group on a benzene ring is benzenecarbaldehyde. However, the common name benzaldehyde is also accepted by IUPAC. Other aromatic aldehydes are hence named as substituted benzaldehydes.

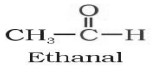




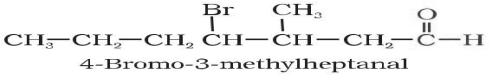
4-Nitrobenzenecarbaldehyde
or
4-Nitrobenzaldehyde

Propane–1,2,3-tricarbaldehyde

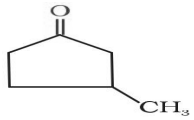
[Note: To give identical treatment to all aldehydic groups, the compound is named as shown above.]



Ethanal

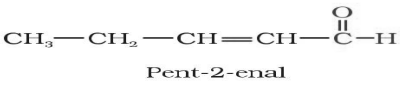


4-Bromo-3-methylheptanal

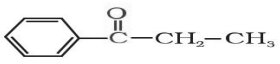


3-Methylcyclopentanone

Cyclohexanecarbaldehyde

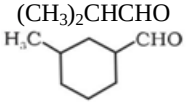
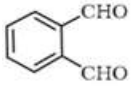
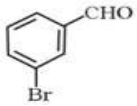
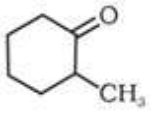


Pent-2-enal



1-Phenylpropan-1-one

The common and IUPAC names of some aldehydes and ketones are given in below table.

Structure	Common name	IUPAC name
Aldehydes		
HCHO	Formaldehyde	Methanal
CH_3CHO	Acetaldehyde	Ethanal
$(CH_3)_2CHCHO$	Isobutyraldehyde	2-Methylpropanal
	γ-Methylcyclohexane	3-Methylcyclohexanecarbaldehyde
$CH_3CH(OCH_3)CHO$	α-Methoxypropionaldehyde	2-Methoxypropanal
$CH_3CH_2CH_2CH_2CHO$	Valeraldehyde	Pentanal
$CH_2=CHCHO$	Acrolein	Prop-2-enal
	Phthaldehyde	Benzene-1,2-dicarbaldehyde
	*m*-Bromobenzaldehyde	3-Bromobenzenecarbaldehyde or 3-Bromobenzaldehyde
Ketones		
$CH_3COCH_2CH_2CH_3$	Methyl *n*-propyl ketone	Pentan-2-one
$(CH_3)_2CHCOCH(CH_3)_2$	Diisopropyl ketone	2,4-Dimethylpentan-3-one
	α-Methylcyclohexanone	2-Methylcyclohexanone
$(CH_3)_2C=CHCOCH_3$	Mesityl oxide	4-Methylpent-3-en-2-one

Structure of the Carbonyl Group

The carbonyl carbon atom is sp^2-hybridised and forms three sigma (σ) bonds. The fourth valence electron of carbon remains in its *p*-orbital and forms a π-bond with oxygen by overlap with *p*-orbital of an oxygen In addition, the oxygen atom also has two non bonding electron pairs. Thus, the carbonyl carbon and the three atoms attached to it lie in the same plane and the π-electron cloud is above and below this plane. The bond angles are approximately 120° as expected of a trigonal coplanar structure (Fig. 12.1).

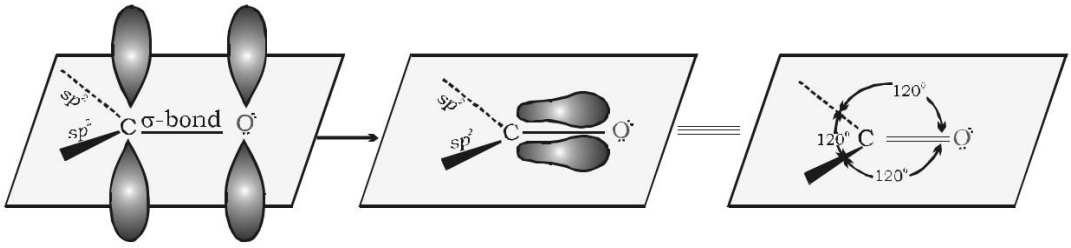


Fig.12.1 Orbital diagram for the formation of carbonyl group

The carbon-oxygen double bond is polarised due to higher electronegativity of oxygen relative to carbon. Hence, the carbonyl – carbon is an electrophilic (Lewis acid), and carbonyl oxygen, a nucleophilic (Lewis base) centre.Carbonyl compounds have substantial dipole moments and are polar than ethers. The high polarity of the carbonyl group is explained on the basis of resonance involving a neutral (A) and a dipolar (B) structures as shown.

12.12 Preparation of Aldehydes and Ketones

Some important methods for the preparation of aldehydes and ketones are as follows:

Preparation of Aldehydes .

(iii) By oxidation of alcohols

Aldehydes are generally prepared by oxidation of primary alcohols .

$$RCH_2OH + (O) \xrightarrow{CrO_3} RCHO$$

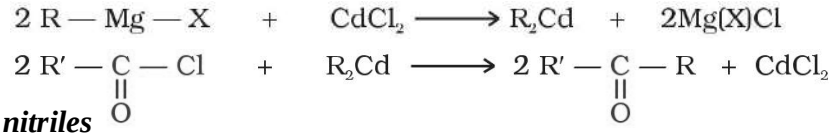
(iv) ***By dehydrogenation of alcohols***

This method is suitable for volatile alcohols and is of industrial application. In this method alcohol vapours are passed over heavy metal catalysts (Ag or Cu). Primary and secondary alcohols give aldehydes and ketones, respectively (Unit 11, Class XII).

Preparation of Ketones

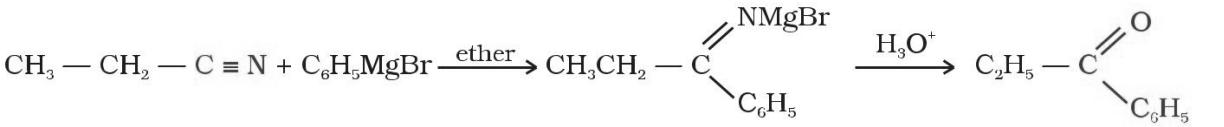
From acyl chlorides

Treatment of acyl chlorides with dialkylcadmium, prepared by the reaction of cadmium chloride with Grignard reagent, gives ketones.



(i) From nitriles

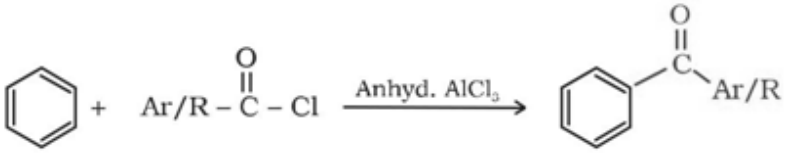
Treating a nitrile with Grignard reagent followed by hydrolysis yields a ketone.



Propiophenone
(1-Phenylpropanone)

3. From benzene or substituted benzenes

When benzene or substituted benzene is treated with acid chloride in the presence of anhydrous aluminium chloride, it affords the corresponding ketone. This reaction is known as **Friedel-Crafts acylation reaction**.



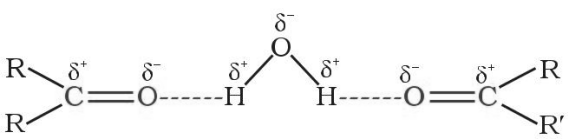
Physical Properties

The physical properties of aldehydes and ketones are described as follows.

Methanal is a gas at room temperature. Ethanal is a volatile liquid. Other aldehydes and ketones are liquid or solid at room temperature. The boiling points of aldehydes and ketones are higher than hydrocarbons and ethers of comparable molecular masses. It is due to weak molecular association in aldehydes and ketones arising out of the dipole-dipole interactions. Also, their boiling points are lower than those of alcohols of similar molecular masses due to absence of intermolecular hydrogen bonding. The following compounds of molecular masses 58 and 60 are ranked in order of increasing boiling points.

	b.p.(K)	**Molecular Mass**
n-Butane	273	58
Methoxyethane	281	60
Propanal	322	58
Acetone	329	58
Propan-1-ol	370	60

The lower members of aldehydes and ketones such as methanal, ethanal and propanone are miscible with water in all proportions, because they form hydrogen bond with water.



However, the solubility of aldehydes and ketones decreases rapidly on increasing the length of alkyl chain. All aldehydes and ketones are fairly soluble in organic solvents like benzene, ether, methanol, chloroform, etc. The lower aldehydes have sharp pungent odours. As the size of the molecule increases, the odour becomes less pungent and more fragrant. In fact, many naturally occurring aldehydes and ketones are used in the blending of perfumes and flavouring agents.

12.13 Chemical Reactions of Aldehydes and Ketones

Since aldehydes and ketones both possess the carbonyl functional group, they undergo similar chemical reactions.

(iii) Nucleophilic addition reactions

Contrary to electrophilic addition reactions observed in alkenes (refer Unit 13, Class XI), the aldehydes and ketones undergo nucleophilic addition reactions.

Mechanism of nucleophilic addition reactions

A nucleophile attacks the electrophilic carbon atom of the polar carbonyl group from a direction approximately perpendicular to the plane of sp^2 hybridised orbitals of carbonyl carbon (Fig. 12.2). The hybridisation of carbon changes from sp^2 to sp^3 in this process, and a tetrahedral alkoxide intermediate is produced. This intermediate captures a proton from the reaction medium to give the The net result is addition of electrically neutral product. The net result is addition of Nu^- and H^+ across the carbon oxygen double bond

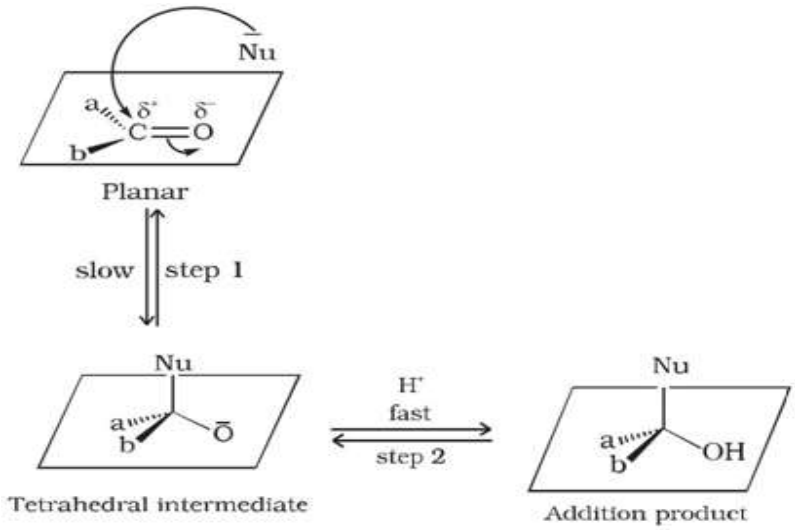


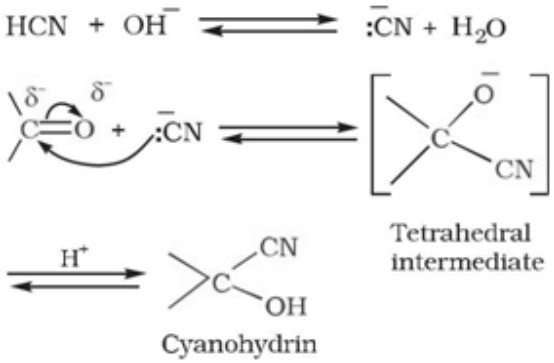
Fig.12.2 Nucleophilic attack on carbonyl carbon

Reactivity

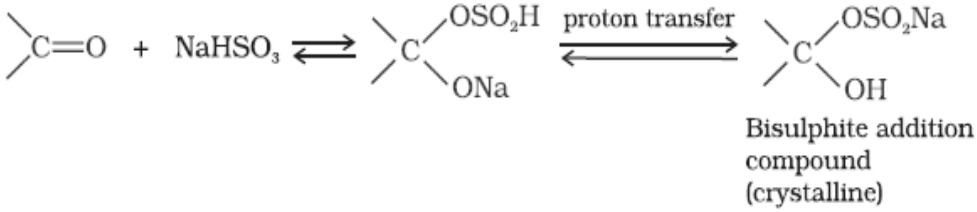
Aldehydes are generally more reactive than ketones in nucleophilic addition reactions due to steric and electronic reasons. Sterically, the presence of two relatively large substituents in ketones hinders the approach of nucleophile to carbonyl carbon than in aldehydes having only one such substituent. Electronically, aldehydes are more reactive than ketones because two alkyl groups reduce the electrophilicity of the carbonyl more effectively than in former.

Some important examples of nucleophilic addition and nucleophilic addition-elimination reactions:

(a) *Addition of hydrogen cyanide (HCN)*: Aldehydes and ketones react with hydrogen cyanide (HCN) to yield cyanohydrins. This reaction occurs very slowly with pure HCN. Therefore, it is catalysed by a base and the generated cyanide ion (CN^-) being a stronger nucleophile readily adds to carbonyl compounds to yield corresponding cyanohydrins. Cyanohydrins are useful synthetic intermediates.

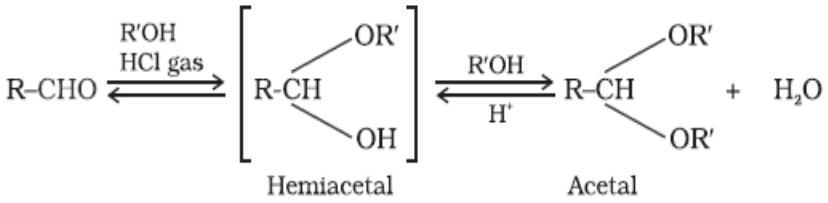


(b) *Addition of sodium hydrogensulphite:* Sodium hydrogensulphite adds to aldehydes and ketones to form the addition products. The position of the equilibrium lies largely to the right hand side for most aldehydes and to the left for most ketones due to steric reasons. The hydrogensulphite addition compound is water soluble and can be converted back to the original carbonyl compound by treating it with dilute mineral acid or alkali. Therefore, these are useful for separation and purification of aldehydes.

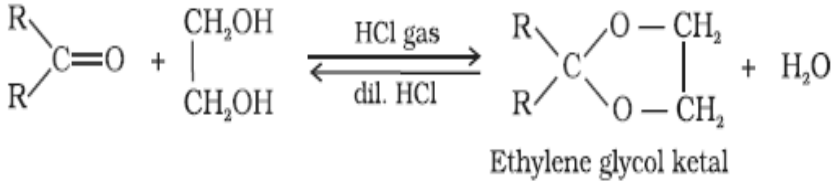


(c) *Addition of Grignard reagents:* (refer Unit 11, Class XII).

(d) *Addition of alcohols:* Aldehydes react with one equivalent of monohydric alcohol in the presence of dry hydrogen chloride to yield alkoxyalcohol intermediate, known as hemiacetals,which further react with one more molecule of alcohol to give *a gem*-dialkoxy compound known as acetal as shown in the reaction. Ketones react with ethylene glycol under similar conditions to form cyclic products known as ethylene glycol ketals.

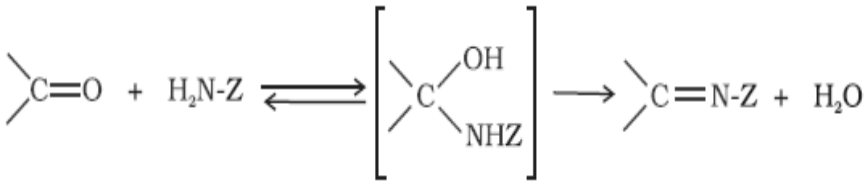


Dry hydrogen chloride protonates the oxygen of the carbonyl compounds and therefore, increases the electrophilicity of the carbonyl carbon facilitating the nucleophilic attack of ethylene glycol. Acetals and ketals are hydrolysed with aqueous mineral acids to yield corresponding aldehydes and ketones respectively.

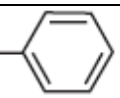


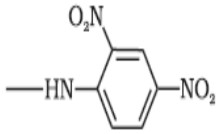
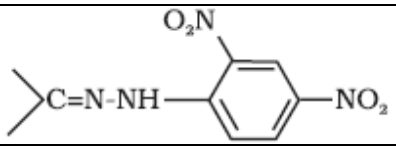
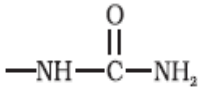
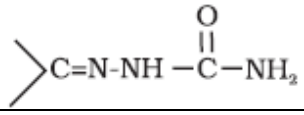
(e) *Addition of ammonia and its derivatives*: Nucleophiles, such as ammonia and its derivatives H2N-Z add to the carbonyl group of aldehydes and ketones. The reaction is reversible and catalysed by acid. The equilibrium favours the product formation due to rapid dehydration of the intermediate to form >C=N-Z.

Z = Alkyl, aryl, OH, NH2, C6H5NH, NHCONH2, etc.



Some N-Substituted Derivatives of Aldehydes and Ketones (>C=N-Z)

Z	Reagent name	Carbonyl derivative	Product name
-H	Ammonia	>C=NH	Imine
-R	Amine	>C=NR	Substituted imine (Schiff's base)
—OH	Hydroxylamine	>C=N–OH	Oxime
—NH2	Hydrazine	>C=N–NH₂	Hydrazone
—HN– 	Phenylhydrazine	>C=N-NH– 	Phenylhydrazone

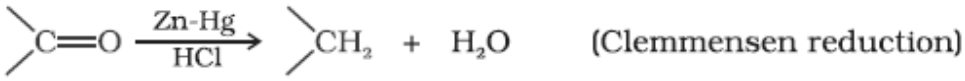
	2,4-Dinitrophenyl-hydrazine		2,4 Dinitrophenyl-hydrazone
	Semicarbazide		Semicarbazone

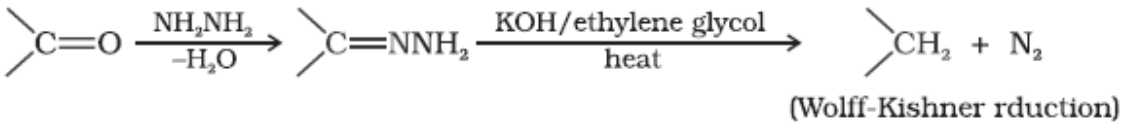
** 2,4-DNP-derivatives are yellow, orange or red solids, useful for characterisation of aldehydes and ketones.*

2. Reduction

(i) *Reduction to alcohols:* Aldehydes and ketones are reduced to primary and secondary alcohols respectively by sodium borohydride (NaBH4) or lithium aluminium hydride (LiAlH4) as well as by catalytic hydrogenation (Unit 11, Class XII).

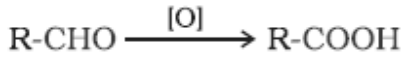
(ii) *Reduction to hydrocarbons:* The carbonyl group of aldehydes and ketones is reduced to CH2 group on treatment with zincamalgam and concentrated hydrochloric acid [**Clemmensen reduction**] or with hydrazine followed by heating with sodium or potassium hydroxide in high boiling solvent such as ethylene glycol (**Wolff-Kishner reduction**).



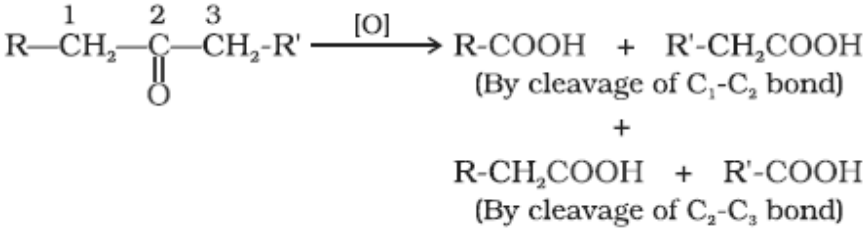


3. Oxidation

Aldehydes differ from ketones in their oxidation reactions. Aldehydes are easily oxidised to carboxylic acids on treatment with common oxidising agents like nitric acid, potassium permanganate, potassium dichromate, etc. Even mild oxidising agents, mainly Tollens' reagent and Fehlings' reagent also oxidise aldehydes.



Ketones are generally oxidised under vigorous conditions, i.e., strong oxidising agents and at elevated temperatures. Their oxidation involves carbon-carbon bond cleavage to afford a mixture of carboxylic acids having lesser number of carbon atoms than the parent ketone.



The mild oxidising agents given below are used to distinguish aldehydes from ketones:

(i) Tollens' test: On warming an aldehyde with freshly prepared ammoniacal silver nitrate solution (Tollens' reagent), a bright silver mirror is produced due to the formation of silver metal. The aldehydes are oxidised to corresponding carboxylate anion. The reaction occurs in alkaline medium.

$$RCHO + 2[Ag(NH_3)_2]^+ + 3\,\bar{O}H \longrightarrow RCO\bar{O} + 2Ag + 2H_2O + 4NH_3$$

(ii) Fehling's test: Fehling reagent comprises of two solutions, Fehling solution A and Fehling solution B. Fehling solution A is aqueous copper sulphate and Fehling solution B is alkaline sodium potassium tartarate (Rochelle salt). These two solutions are mixed in equal amounts before test. On heating an aldehyde with Fehling's reagent, a reddish brown precipitate is obtained. Aldehydes are oxidised to corresponding carboxylate anion. Aromatic aldehydes do not respond to this test.

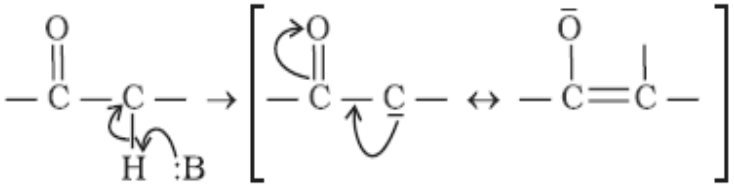
$$R\text{-}CHO + 2Cu^{2+} + 5\bar{O}H \longrightarrow \underset{\text{Red-brown ppt}}{RCO\bar{O} + Cu_2O} + 3H_2O$$

(iii) Oxidation of methyl ketones by haloform reaction: Aldehydes and ketones having at least one methyl group linked to the carbonyl carbon atom (methyl ketones) are oxidised by sodium hypohalite to sodium salts of corresponding carboxylic acids having one carbon atom less than that of carbonyl compound. The methyl group is converted to haloform. This oxidation does not affect a carbon-carbon double bond, if present in the molecule. Iodoform reaction with sodium hypoiodite is also used for detection of CH3CO group or CH3CH(OH) group which produces CH3CO group on oxidation.

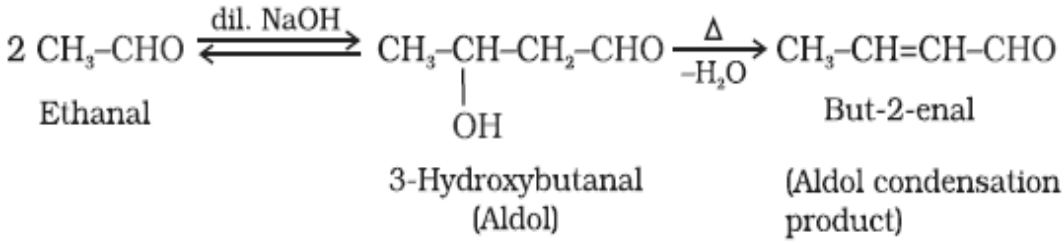
4. Reactions due to a-hydrogen

Acidity of α-hydrogens of aldehydes and ketones: The aldehydes and ketones undergo a number of reactions due to the acidic nature of α-hydrogen.

The acidity of α-hydrogen atoms of carbonyl compounds is due to the strong electron withdrawing effect of the carbonyl group and resonance stabilisation of the conjugate base.

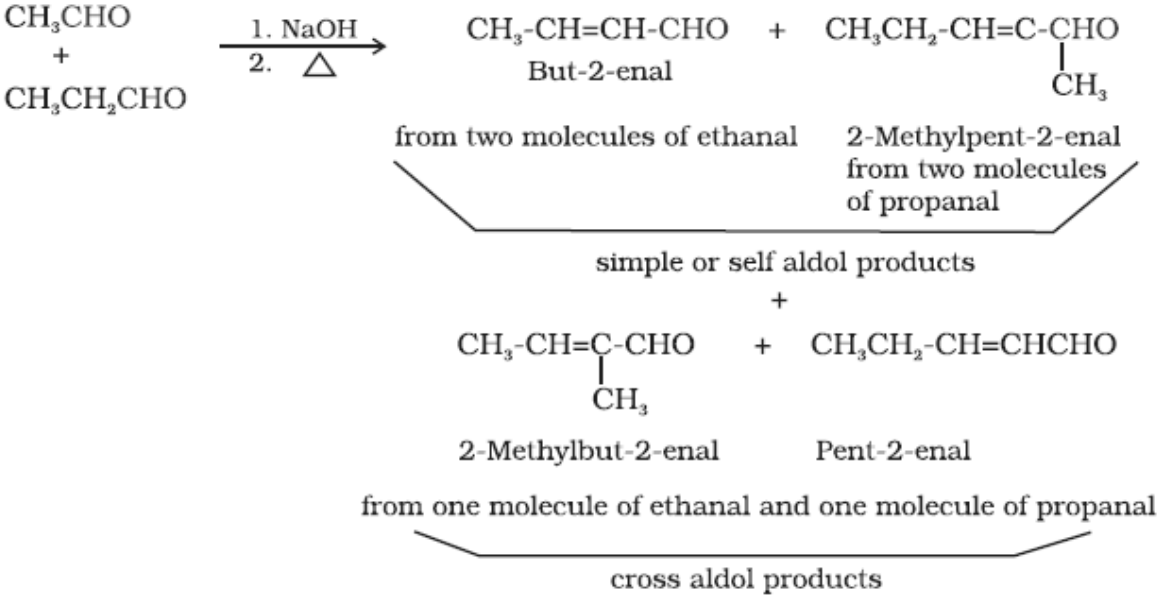


(i) Aldol condensation: Aldehydes and ketones having at least one α-hydrogen undergo a reaction in the presence of dilute alkali as catalyst to form β-hydroxy aldehydes (aldol) or β-hydroxy ketones (ketol), respectively. This is known as **Aldol reaction.**

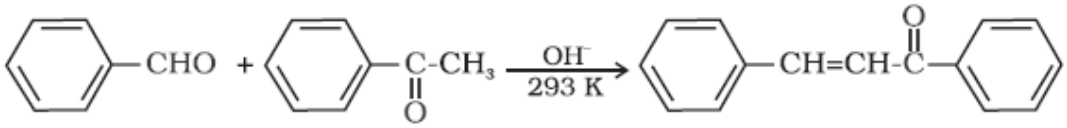


The name aldol is derived from the names of the two functional groups, aldehyde and alcohol, present in the products. The aldol and ketol readily lose water to give α,β-unsaturated carbonyl compounds which are aldol condensation products and the reaction is called **Aldol condensation.** Though ketones give ketols (compounds containing a keto and alcohol groups), the general name aldol condensation still applies to the reactions of ketones due to their similarity with aldehydes.

(ii) Cross aldol condensation: When aldol condensation is carried out between two different aldehydes and / or ketones, it is called **cross aldol condensation**. If both of them contain α-hydrogen atoms, it gives a mixture of four products. This is illustrated below by aldol reaction of a mixture of ethanal and propanal.



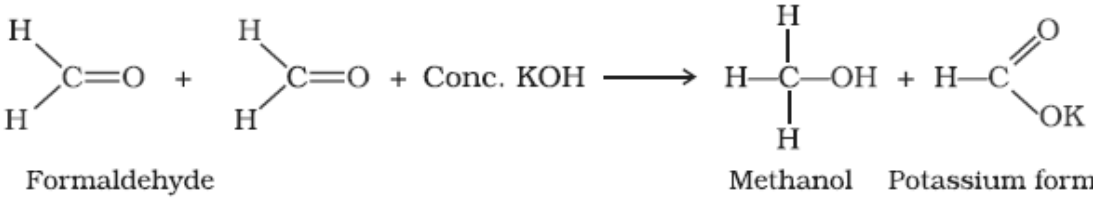
Ketones can also be used as one component in the cross aldol reactions.



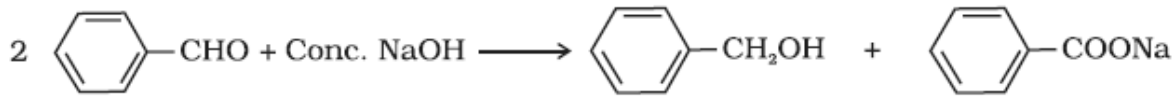
1, 3-Diphenylprop-2-en-1-one
(Benzalacetophenone)
(Major product)

5. Other reactions

(i) Cannizzaro reaction: Aldehydes which do not have an α-hydrogen atom, undergo self. oxidation and reduction (disproportionation) reaction on treatment with concentrated alkali. In this reaction, one molecule of the aldehyde is reduced to alcohol while another is oxidised to carboxylic acid salt.



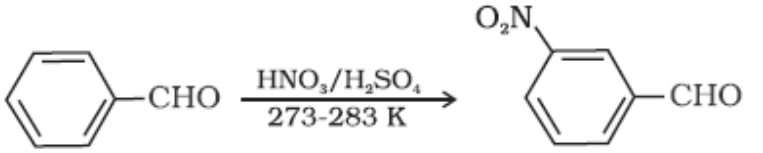
Formaldehyde Methanol Potassium formate



Benzaldehyde Benzyl alcohol Sodium benzoate

-

(ii) Electrophilic substitution reaction: Aromatic aldehydes and ketones undergo electrophilic substitution at the ring in which the carbonyl group acts as a deactivating and *meta*-directing group.



Benzaldehyde *m*-Nitrobenzaldehyde

12.14 Uses of Aldehydes and Ketones

In chemical industry aldehydes and ketones are used as solvents, starting materials and reagents for the synthesis of other products. Formaldehyde is well known as formalin (40%) solution used to preserve biological specimens and to prepare bakelite (a phenol-formaldehyde resin), urea-formaldehyde glues and other polymeric products. Acetaldehyde is used primarily as a starting material in the manufacture of acetic acid, ethyl acetate, vinyl acetate, polymers and drugs. Benzaldehyde is used in perfumery and in dye industries. Acetone and ethyl methyl ketone are common industrial solvents. Many aldehydes and ketones, e.g., butyraldehyde, vanillin, acetophenone, camphor, etc.are well known for their odours and flavours.

CARBOXYLIC ACIDS

Carbon compounds containing a carboxyl functional group, –COOH are called carboxylic acids. The carboxyl group, consists of a *carbonyl* group attached to a *hydroxyl* group, hence its name *carboxyl*. Carboxylic acids may be aliphatic (RCOOH) or aromatic (ArCOOH) depending on the group, alkyl or aryl, attached to carboxylic carbon. Large number of carboxylic acids are found in nature. Some higher members of aliphatic carboxylic acids (C12 – C18) known as **fatty acids,** occur in natural fats as esters of glycerol. Carboxylic acids serve as starting material for several other important organic compounds such as anhydrides, esters, acid chlorides, amides, etc.

12.15 Nomenclature and Structure of Carboxyl Group

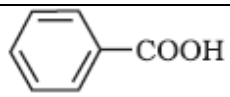
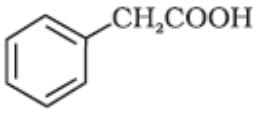
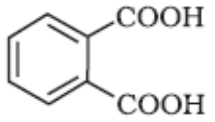
Nomenclature

Since carboxylic acids are amongst the earliest organic compounds to be isolated from nature, a large number of them are known by their common names. The common names end with the suffix –*ic* acid and have been derived from Latin or Greek names of their natural sources. For example, formic acid (HCOOH) was first obtained from red ants (Latin: *formica* means ant), acetic acid (CH3COOH) from vinegar (Latin: *acetum*, means vinegar), butyric acid (CH3CH2CH2COOH) from rancid butter (Latin: *butyrum*, means butter).

In the IUPAC system, aliphatic carboxylic acids are named by replacing the ending –*e* in the name of the corresponding alkane with – *oic acid*. In numbering the carbon chain, the carboxylic carbon is numbered one. For naming compounds containing more than one carboxyl group, the ending –*e* of the alkane is retained. The number of carboxyl groups are indicated by adding the multiplicative prefix, *di, tri,* etc. to the term **oic**. The position of –COOH groups are indicated by the arabic numeral before the multiplicative prefix. Some of the carboxylic acids along with their common and IUPAC names are listed in below Table.

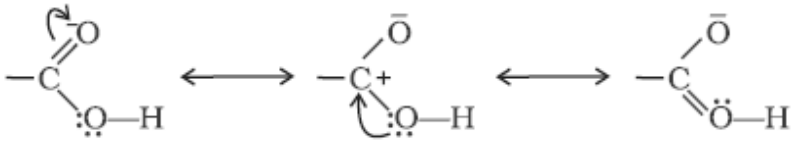
Names and Structures of Some Carboxylic Acids

Structure	Common name	IUPAC name
HCOOH	Formic acid	Methanoic acid
CH_3COOH	Acetic acid	Ethanoic acid
CH_3CH_2COOH	Propionic acid	Propanoic acid
$CH_3CH_2CH_2COOH$	Butyric acid	Butanoic acid
$(CH_3)_2CHCOOH$	Isobutyric acid	2-Methylpropanoic acid
HOOC-COOH	Oxalic acid	Ethanedioic acid
HOOC -CH_2-COOH	Malonic acid	Propanedioic acid
HOOC -$(CH_2)_2$-COOH	Succinic acid	Butanedioic acid
HOOC -$(CH_2)_3$-COOH	Glutaric acid	Pentanedioic acid
HOOC -$(CH_2)_4$-COOH	Adipic acid	Hexanedioic acid
HOOC -CH_2-CH(COOH)-CH_2-COOH	–	Propane-1, 2, 3-tricarboxylic acid

	Benzoic acid	Benzenecarboxylic acid (Benzoic acid)
	Phenylacetic acid	2-Phenylethanoic acid
	Phthalic acid	Benzene-1, 2-dicarboxylic acid

Structure of Carboxyl Group

In carboxylic acids, the bonds to the carboxyl carbon lie in one plane and are separated by about 120°. The carboxylic carbon is less electrophilic than carbonyl carbon because of the possible resonance structure shown below:



12.16 Methods of Preparation of Carboxylic Acids

Some important methods of preparation of carboxylic acids are as follows.

1. From primary alcohols and aldehydes

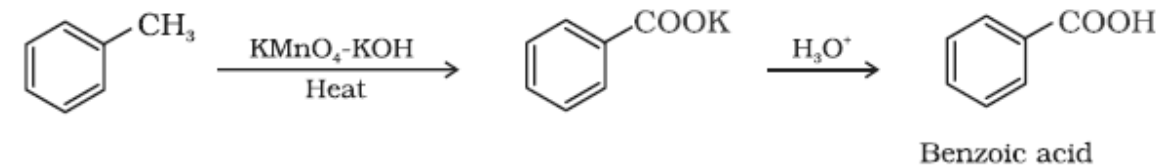
Primary alcohols are readily oxidised to carboxylic acids with common oxidising agents such as potassium permanganate (KMnO4) in neutral, acidic or alkaline media or by potassium dichromate (K2Cr2O7) and chromium trioxide (CrO3) in acidic media Carboxylic acids are also prepared from aldehydes by the use of mild oxidising agents (Section 12.4).

$$RCH_2OH \xrightarrow[2.\ H_3O^+]{1.\ \text{alkaline } KMnO_4} RCOOH$$

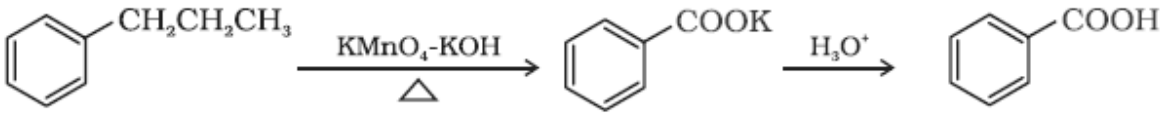
$$\underset{\text{1-Decanol}}{CH_3(CH_2)_8CH_2OH} \xrightarrow{CrO_3\text{-}H_2SO_4} \underset{\text{Decanoic acid}}{CH_3(CH_2)_8COOH}$$

2. From alkylbenzenes

Aromatic carboxylic acids can be prepared by vigorous oxidation of alkyl benzenes with chromic acid or acidic or alkaline potassium permanganate. The entire side chain is oxidised to the carboxyl group irrespective of length of the side chain. Primary and secondary alkyl groups are oxidised in this manner while tertiary group is not affected. Suitably substituted alkenes are also oxidised to carboxylic acids with these oxidising reagents (refer Unit 13, Class XI).



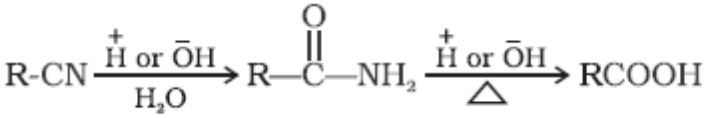
Benzoic acid

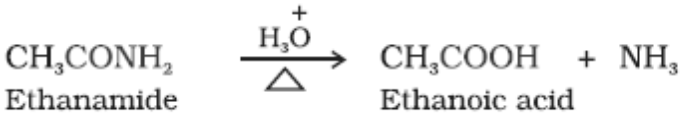


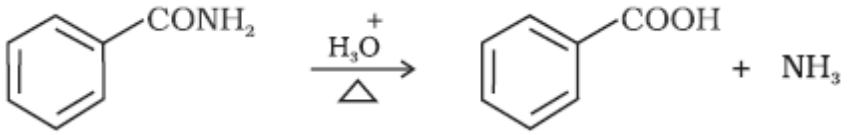
Benzoic acid

3. From nitriles and amides

Nitriles are hydrolysed to amides and then to acids in the presence of H^+ or OH as catalyst. Mild reaction conditions are used to stop the reaction at the amide stage.



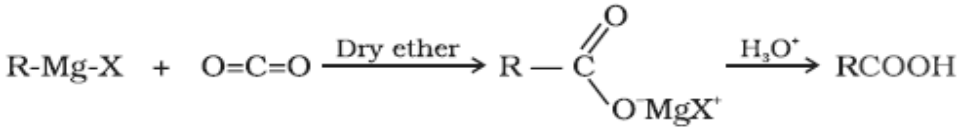




Benzamide Benzoic acid

4. From Grignard reagents

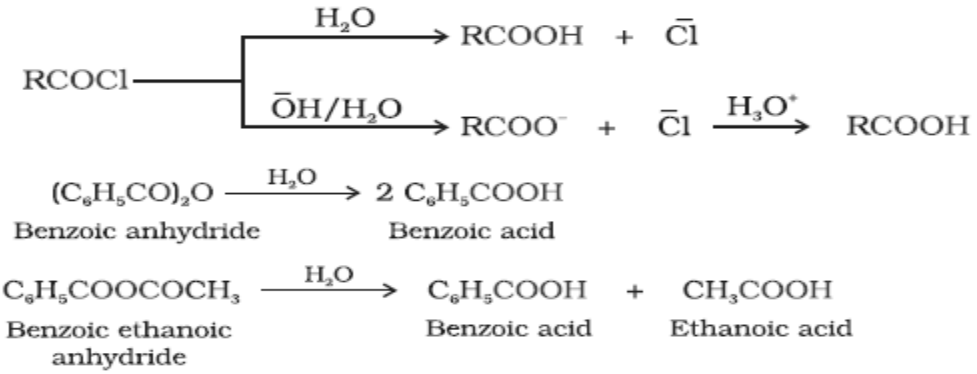
Grignard reagents react with carbon dioxide (dry ice) to form salts of carboxylic acids which in turn give corresponding carboxylic acids after acidification with mineral acid.



As we know, the Grignard reagents and nitriles can be prepared from alkyl halides (refer Unit 10, Class XII). The above methods (3 and 4) are useful for converting alkyl halides into corresponding carboxylic acids having one carbon atom more than that present in alkyl halides (ascending the series).

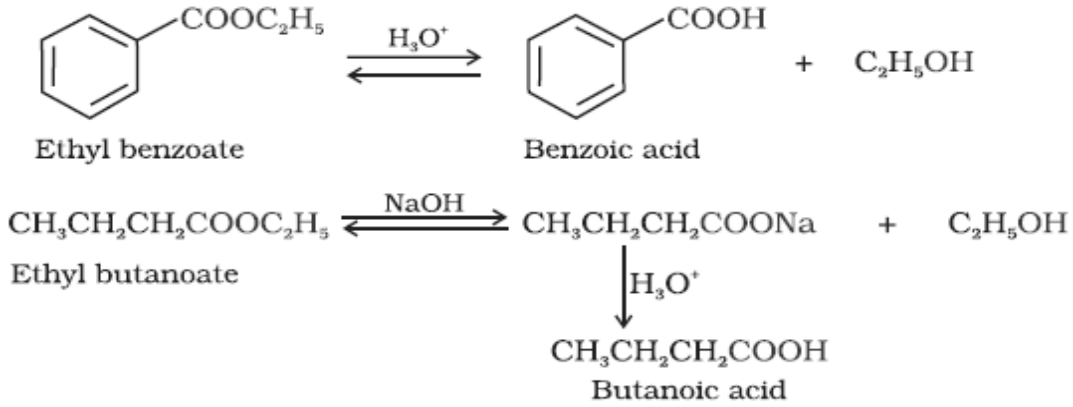
5. From acyl halides and anhydrides

Acid chlorides when hydrolysed with water give carboxylic acids or more readily hydrolysed with aqueous base to give carboxylate ions which on acidification provide corresponding carboxylic acids. Anhydrides on the other hand are hydrolysed to corresponding acid(s) with water.



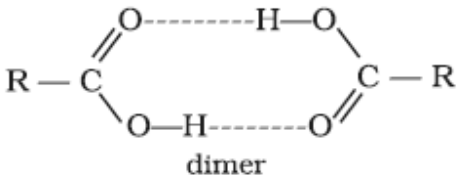
6. From esters

Acidic hydrolysis of esters gives directly carboxylic acids while basic hydrolysis gives carboxylates, which on acidification give corresponding carboxylic acids.



12.14 Physical Properties

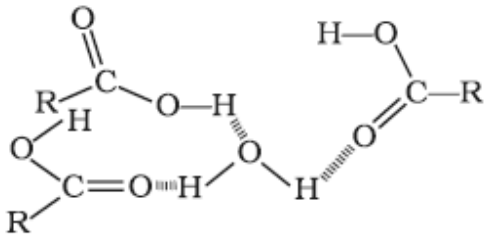
Aliphatic carboxylic acids upto nine carbon atoms are colourless liquids at room temperature with unpleasant odours. The higher acids are wax like solids and are practically odourless due to their low volatility. Carboxylic acids are higher boiling liquids than aldehydes, ketones and even alcohols of comparable molecular masses. This is due to more extensive association of carboxylic acid molecules through intermolecular hydrogen bonding. The hydrogen bonds are not broken completely even in the vapour phase. In fact, most carboxylic acids exist as dimer in the vapour phase or in the aprotic solvents.



In vapour state or in aprotic solvent

Simple aliphatic carboxylic acids having upto four carbon atoms are miscible in water due to the formation of hydrogen bonds with water. The solubility decreases with increasing number of carbon atoms. Higher carboxylic acids are practically insoluble in water due to the

increased hydrophobic interaction of hydrocarbon part. Benzoic acid, the simplest aromatic carboxylic acid is nearly insoluble in cold water. Carboxylic acids are also soluble in less polar organic solvents like benzene, ether, alcohol, chloroform, etc.



Hydrogen bonding of RCOOH with H_2O

12.18 Chemical Reactions

The reaction of carboxylic acids are classified as follows:

Reactions Involving Cleavage of O–H Bond

Acidity

Reactions with metals and alkalies

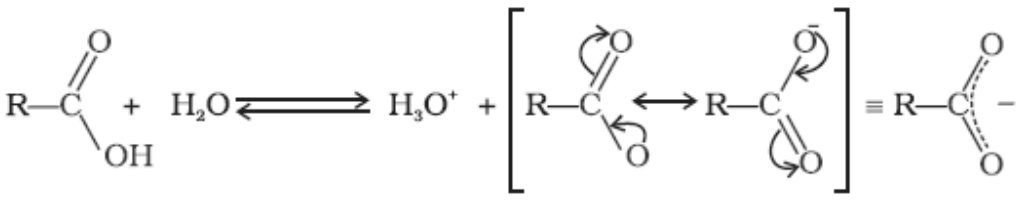
The carboxylic acids like alcohols evolve hydrogen with electropositive metals and form salts with alkalies similar to phenols. However, unlike phenols they react with weaker bases such as carbonates and hydrogencarbonates to evolve carbon dioxide. This reaction is used to detect the presence of carboxyl group in an organic

$$2R\text{-}COOH + 2Na \longrightarrow \underset{\text{Sodium carboxylate}}{2R\text{-}CO\bar{O}Na^+} + H_2$$

$$R\text{-}COOH + NaOH \longrightarrow R\text{-}CO\bar{O}Na^+ + H_2O$$

$$R\text{-}COOH + NaHCO_3 \longrightarrow R\text{-}CO\bar{O}Na^+ + H_2O + CO_2$$

Carboxylic acids dissociate in water to give resonance stabilised carboxylate anions and hydronium ion.



For the above reaction:

$$K_{eq} = \frac{[H_3\overset{+}{O}]\,[RCO\bar{O}]}{[H_2O]\,[RCOOH]} \qquad K_a = K_{eq}\,[H_2O] = \frac{[H_3\overset{+}{O}]\,[RCO\bar{O}]}{[RCOOH]}$$

where *Keq*, is equilibrium constant and *Ka* is the acid dissociation constant.

For convenience, the strength of an acid is generally indicated by its p*Ka* value rather than its *Ka* value.

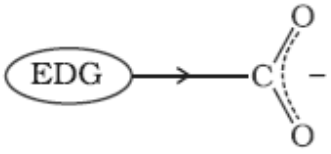
$$pKa = -\log Ka$$

The p*Ka* of hydrochloric acid is –7.0, where as p*Ka* of trifluoroacetic acid (the strongest organic acid), benzoic acid and acetic acid are 0.23, 4.19 and 4.76, respectively. Smaller the

p*Ka*, the stronger the acid (the better it is as a proton donor). Strong acids have p*Ka* values < 1, the acids with p*Ka* values between 1 and 5 are considered to be moderately strong acids, weak acids have p*Ka* values between 5 and 15, and extremely weak acids have p*Ka* values >15. Carboxylic acids are weaker than mineral acids, but they are stronger acids than alcohols and many simple phenols (p*Ka* is ~16 for ethanol and 10 for phenol). In fact, carboxylic acids are amongst the most acidic organic compounds you have studied so far. You already know why phenols are more acidic than alcohols. The higher acidity of carboxylic acids as compared to phenols can be understood similarly. The conjugate base of carboxylic acid, a carboxylate ion, is stabilised by two equivalent resonance structures in which the negative charge is at the more electronegative oxygen atom. The conjugate base of phenol, a phenoxide ion, has non-equivalent resonance structures in which the negative charge is at the less electronegative carbon atom. Therefore, resonance in phenoxide ion is not as important as it is in carboxylate ion. Further, the negative charge is delocalised over two electronegative oxygen atoms in carboxylate ion whereas it is less effectively delocalised over one oxygen atom and less electronegative carbon atoms in phenoxide ion (Unit 11, Class XII). Thus, the carboxylate ion is more stabilised than phenoxide ion, so carboxylic acids are more acidic than phenols.

Effect of substituents on the acidity of carboxylic acids: Substituents may affect the stability of the conjugate base and thus, also affect the acidity of the carboxylic acids. Electron withdrawing groups increase the acidity of carboxylic acids by stabilising the conjugate base through delocalisation of the negative charge by inductive and/or resonance effects. Conversely, electron donating groups decrease the acidity by destabilising the conjugate base.

Electron withdrawing group (EWG) stabilises the carboxylate anion and strengthens the acid



Electron donating group (EDG) destabilises the carboxylate anion and weakens the acid

The effect of the following groups in increasing acidity order is

Ph < I < Br < Cl < F < CN < NO2 < CF3

Thus, the following acids are arranged in order of decreasing acidity (based on p*Ka* values):

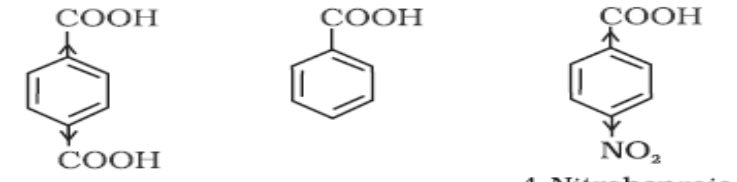
$CF_3COOH > CCl_3COOH > CHCl_2COOH > NO_2CH_2COOH > NC\text{-}CH_2COOH >$ ←

$FCH_2COOH > ClCH_2COOH > BrCH_2COOH > HCOOH > ClCH_2CH_2COOH >$
(continue) ←

$C_6H_5COOH > C_6H_5CH_2COOH > CH_3COOH > CH_3CH_2COOH$
(continue) ←

Direct attachment of groups such as phenyl or vinyl to the carboxylic acid, increases the acidity of corresponding carboxylic acid, contrary to the decrease expected due to resonance effect shown below:

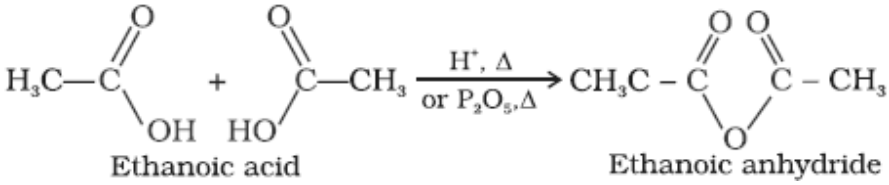
This is because of greater electronegativity of *sp2* hybridised carbon to which carboxyl carbon is attached. The presence of electron withdrawing group on the phenyl of aromatic carboxylic acid increases their acidity while electron donating groups decrease their acidity.



Reactions Involving Cleavage of C–OH Bond

1. Formation of anhydride

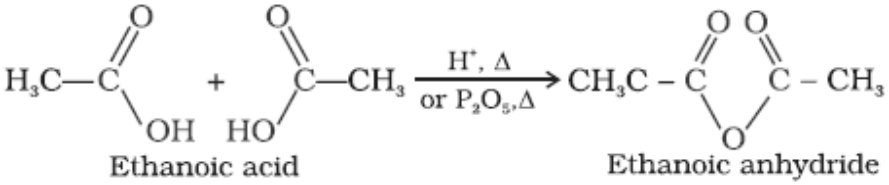
Carboxylic acids on heating with mineral acids such as H2SO4 or with P2O5 give corresponding anhydride.



Reactions Involving Cleavage of C–OH Bond

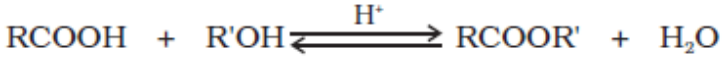
1. Formation of anhydride

Carboxylic acids on heating with mineral acids such as H2SO4 or with P2O5 give corresponding anhydride.



2. Esterification

Carboxylic acids are esterified with alcohols or phenols in the presence of a mineral acid such as concentrated H2SO4 or HCl gas as a catalyst.



3. Reactions with PCl_5, PCl_3 and $SOCl_2$

The hydroxyl group of carboxylic acids, behaves like that of alcohols and is easily replaced by chlorine atom on treating with PCl5, PCl3 or SOCl2. Thionyl chloride (SOCl2) is preferred because the other two products are gaseous and escape the reaction mixture making the purification of the products easier.

$$RCOOH + PCl_5 \longrightarrow RCOCl + PCl_3 + HCl$$

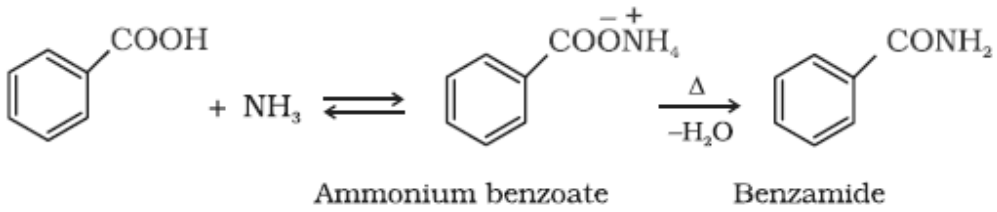
$$3RCOOH + PCl_3 \longrightarrow 3RCOCl + H_3PO_3$$

$$RCOOH + SOCl_2 \longrightarrow RCOCl + SO_2 + HCl$$

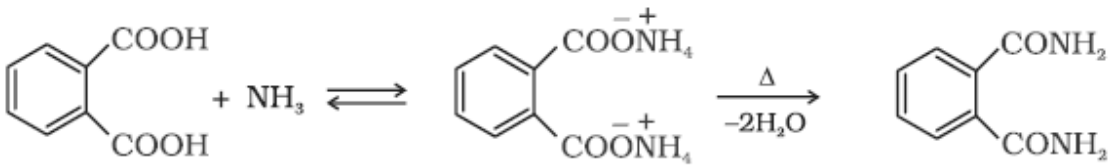
4. Reaction with ammonia

Carboxylic acids react with ammonia to give ammonium salt which on further heating at high temperature give amides. For example:

$$CH_3COOH + NH_3 \rightleftharpoons \underset{\text{Ammonium acetate}}{CH_3COO^-NH_4^+} \xrightarrow[-H_2O]{\Delta} \underset{\text{Acetamide}}{CH_3CONH_2}$$

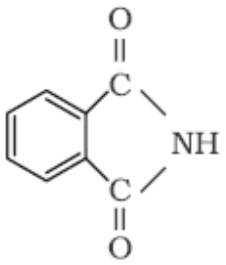


Ammonium benzoate Benzamide



Ammonium phthalate Phthalamide

$$\xrightarrow[\text{Strong heating}]{-NH_3}$$



Phthalimide

Reactions Involving –COOH Group

1. Reduction

Carboxylic acids are reduced to primary alcohols by lithium aluminium hydride or better with diborane. Diborane does not easily reduce functional groups such as ester, nitro, halo, etc. Sodium borohydride does not reduce the carboxyl group.

$$R\text{-}COOH \xrightarrow[(ii)\ H_3O^+]{(i)\ LiAlH_4/\text{ether} \quad \text{or } B_2H_6} R\text{-}CH_2OH$$

2. Decarboxylation

Carboxylic acids lose carbon dioxide to form hydrocarbons when their sodium salts are heated with sodalime (NaOH and CaO in the ratio of 3 : 1). The reaction is known as decarboxylation.

$$R\text{-}COONa \xrightarrow[Heat]{NaOH \ \& \ CaO} R\text{-}H + Na_2CO_3$$

Alkali metal salts of carboxylic acids also undergo decarboxylation on electrolysis of their aqueous solutions and form hydrocarbons having twice the number of carbon atoms present in the alkyl group of the acid. The reaction is known as **Kolbe electrolysis** (Unit 13, Class XI)**.**

Substitution Reactions in the Hydrocarbon Part

1. Halogenation

Carboxylic acids having an α-hydrogen are halogenated at the α-position on treatment with chlorine or bromine in the presence of small amount of red phosphorus to give α-alocarboxylic acids. The reaction is known as **Hell-Volhard-Zelinsky reaction.**

$$R\text{-}CH_2\text{-}COOH \xrightarrow[(ii)\ H_2O]{(i)\ X_2/\text{Red phosphorus}} R\text{-}\underset{\displaystyle X}{\underset{|}{CH}}\text{-}COOH$$

X = Cl, Br

α – Halocarboxylic acid

12.19 Uses of Carboxylic Acids

Methanoic acid is used in rubber, textile, dyeing, leather and electroplating industries. Ethanoic acid is used as solvent and as vinegar in food industry. Hexanedioic acid is used in the manufacture of nylon-6, 6. Esters of benzoic acid are used in perfumery. Sodium benzoate is used as a food preservative. Higher fatty acids are used for the manufacture of soaps and detergents.

Summary

Aldehydes, **ketones** and **carboxylic acids** are some of the important classes of organic compounds containing carbonyl group. These are highly polar molecules. Therefore, they boil at higher temperatures than the hydrocarbons and weakly polar compounds such as ethers of comparable molecular masses. The lower members are more soluble in water because they form hydrogen bonds with water. The higher members, because of large size of hydrophobic chain of carbon atoms, are insoluble in water but soluble in common organic solvents. Aldehydes are prepared by dehydrogenation or controlled oxidation of primary alcohols and controlled or selective reduction of acyl halides. Aromatic aldehydes may also be prepared by oxidation of (i) methylbenzene with chromyl chloride or CrO3 in the presence of acetic anhydride, (ii) formylation of arenes with carbon monoxide and hydrochloric acid in the presence of anhydrous aluminium chloride, and (iii) cuprous chloride or by hydrolysis of benzal chloride. Ketones are prepared by oxidation of secondary alcohols and hydration of alkynes. Ketones are also prepared by reaction of acyl chloride with dialkylcadmium. A good method for the preparation of aromatic ketones is the **Friedel-Crafts acylation** of aromatic hydrocarbons with acyl chlorides or anhydrides. Both aldehydes and ketones can be prepared by ozonolysis of alkenes. Aldehydes and ketones undergo nucleophilic addition reactions onto the carbonyl group with a number of

nucleophiles such as, HCN, NaHSO3, alcohols (or diols), ammonia derivatives, and **Grignard reagents**. The α-hydrogens in aldehydes and ketones are acidic. Therefore, aldehydes and ketones having at least one α-hydrogen, undergo A**ldol condensation** in the presence of a base to give α-hydroxyaldehydes (aldol) and α-hydroxyketones(ketol), respectively. Aldehydes having no α-hydrogen undergo **Cannizzaro reaction** in the presence of concentrated alkali. Aldehydes and ketones are reduced to alcohols with NaBH4, LiAlH4, or by catalytic hydrogenation. The carbonyl group of aldehydes and ketones can be reduced to a methylene group by **Clemmensen reduction** or **Wolff-Kishner reduction**. Aldehydes are easily oxidised to carboxylic acids by mild oxidising reagents such as **Tollens' reagent** and **Fehling's reagent**. These oxidation reactions are used to distinguish aldehydes from ketones. Carboxylic acids are prepared by the oxidation of primary alcohols, aldehydes and alkenes by hydrolysis of nitriles, and by treatment of Grignard reagents with carbon dioxide. Aromatic carboxylic acids are also prepared by side-chain oxidation of alkylbenzenes. Carboxylic acids are considerably more acidic than alcohols and most of simple phenols. Carboxylic acids are reduced to primary alcohols with LiAlH4, or better with diborane in ether solution and also undergo α-halogenation with Cl2 and Br2 in the presence of red phosphorus (**Hell-Volhard Zelinsky reaction**). Methanal, ethanal, propanone, benzaldehyde, formic acid, acetic acid and benzoic acid are highly useful compounds in industry.

Important Questions

1) Explain the following reactions with a suitable examples
   a) Kolbe's reaction.
   b) Reimer-Tiemann reaction
   c) Williamson's ether synthesis
2) Describe the following
   a) Acetylation
   b) Cannizaro reaction
   c) Cross aldol condensation
   d) Decarboxylation.
3) Write any two methods of preparation of alcohols and phenols.
4) Explain the acidic nature of phenols & compare with that of alcohols.
5) Write the products obtained when ethanol reacts with conc.H_2SO_4 at 443K & 413K . Explain with mechanism.

CHAPTER 13

ORGANIC COMPOUNDS CONTAINING NITROGEN

AMINES

Amines constitute an important class of organic compounds derived by replacing one or more hydrogen atoms of ammonia molecule by alkyl/aryl group(s). In nature, they occur among proteins, vitamins, alkaloids and hormones. Synthetic examples include polymers, dyestuffs and drugs. Two biologically active compounds, namely adrenaline and ephedrine, both containing secondary amino group, are used to increase blood pressure. Novocain, a synthetic amino compound, is used as an anaesthetic in dentistry. Benadryl, a well known antihistaminic drug also contains tertiary amino group. Quaternary ammonium salts are used as surfactants. Diazonium salts are intermediates in the preparation of a variety of aromatic compounds including dyes. In this Unit, you will learn about amines and diazonium salts.

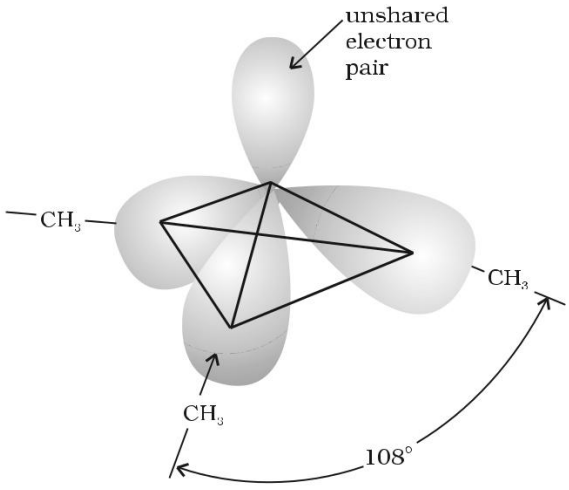
Amines

Amines can be considered as derivatives of ammonia, obtained by replacement of one, two or all the three hydrogen atoms by alkyl and/or aryl groups

For example: $CH_3–NH_2$, $C_6H_5–NH_2$, $CH_3–NH–CH_3$, $CH_3–N(CH_3)_2$

13.1 Structure of amines

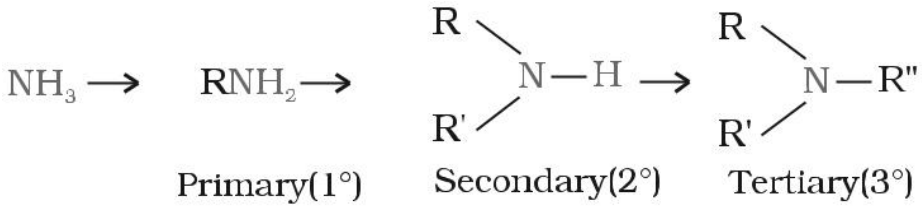
Like ammonia, nitrogen atom of amines is trivalent and carries an unshared pair of electrons. Nitrogen orbitals in amines are therefore, sp3 hybridised and the geometry of amines is pyramidal. Each of the three sp^3 hybridised orbitals of nitrogen overlap with orbitals of hydrogen or carbon depending upon the composition of the amines. The fourth orbital of nitrogen in all amines contains an unshared pair of electrons. Due to the presence of unshared pair of electrons, the angle C–N–E ;(where E is C or H) is less than 109.5 ; for instance, it is 108^o in case of trimethylamine as shown in the follwing figure.



13.2 Classification of amines

Amines are classified as primary (1^o),secondary(2^o),and tertiary (3^o) depending upon the number of hydrogen atoms replaced by alkyl or aryl groups in ammonia molecule. If one hydrogen atom of ammonia is replaced by R or Ar , we get RNH_2 or $ArNH_2$, a primary amine (1^o). If two hydrogen atoms of ammonia or one hydrogen atom of R-NH_2 are replaced

by another alkyl/aryl(R') group, what would you get? You get R-NHR', secondary amine. The second alkyl/aryl group may be same or different. Replacement of another hydrogen atom by alkyl/aryl group leads to the formation of tertiary amine. Amines are said to be 'simple' when all the alkyl or aryl groups are the same, and 'mixed' when they are different.



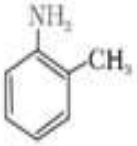
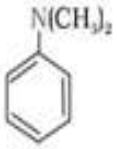
13.3 Nomenclature

In common system,an aliphatic amine is named by prefixing alkyl group to amine,i.e., alkylamine as one word (e.g.,methylamine).In secondary and tertiary amines ,when two or more groups are the same, the prefix di or tri is appended before the name of alkyl group. In IUPAC system, amines are named as alkanamines, derived by replacement of 'e' of alkane by the word amine. For example, CH_3NH_2 is named as methanamine.

In case, more than one amino group is present at different positions in the parent chain, their positions are specified by giving numbers to the carbon atoms bearing $–NH_2$ groups and suitable prefix such as di, tri, etc. is attached to the amine. The letter 'e' of the suffix of the hydrocarbon part is retained. For example, $H_2N–CH_2–CH_2–NH_2$ is named as ethane-1, 2-diamine.In arylamines, $–NH_2$ group is directly attached to the benzenering. $C_6H_5NH_2$ is the simplest example of arylamine.

In common system, it is known as aniline. It is also an accepted IUPAC name. While naming arylamines according to IUPAC system, suffix 'e' of arene is replaced by 'amine'. Thus in IUPAC system, $C_6H_5–NH_2$ is named as benzenamine. Common and IUPAC names of some alkylamines and arylamines are given in Table 13.1.

Table 13.1 Nomenclature of Some Alkylamines and Arylamines

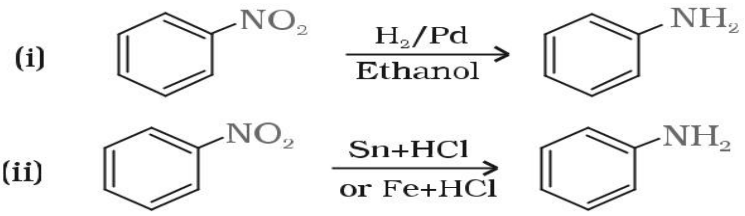
Amine	Common Name	IUPAC Name
$CH_3–CH_2–NH_2$	Ethylamine	Ethanamine
$CH_3–CH_2–CH_2–NH_2$	*n*-Propylamine	Propan-1-amine
$CH_3–\underset{\displaystyle NH_2}{\underset{|}{CH}}–CH_3$	Isopropylamine	Propan-2-amine
$CH_3–\underset{\displaystyle H}{\underset{|}{N}}–CH_2–CH_3$	Ethylmethylamine	N-Methylethanamine
$CH_3–\underset{\displaystyle CH_3}{\underset{|}{N}}–CH_3$	Trimethylamine	N,N-Dimethylmethanamine
$C_2H_5–\underset{\displaystyle C_2H_5}{\underset{|}{N}}–\overset{1}{CH_2}–\overset{2}{CH_2}–\overset{3}{CH_2}–\overset{4}{CH_3}$	*N,N*-Diethylbutylamine	N,N-Diethylbutan-1-amine
$NH_2–\overset{1}{CH_2}–\overset{2}{CH}=\overset{3}{CH_2}$	Allylamine	Prop-2-en-1-amine
$NH_2–(CH_2)_6–NH_2$	Hexamethylenediamine	Hexane-1,6-diamine
	Aniline	Aniline or Benzenamine
	*o*-Toluidine	2-Aminotoluene
	*p*-Bromoaniline	4-Bromobenzenamine or 4-Bromoaniline
	*N,N*-Dimethylaniline	N,N-Dimethylbenzenamine

13.4 Preparation of Amines

Amines are prepared by the following methods:

1. Reduction of nitro compounds

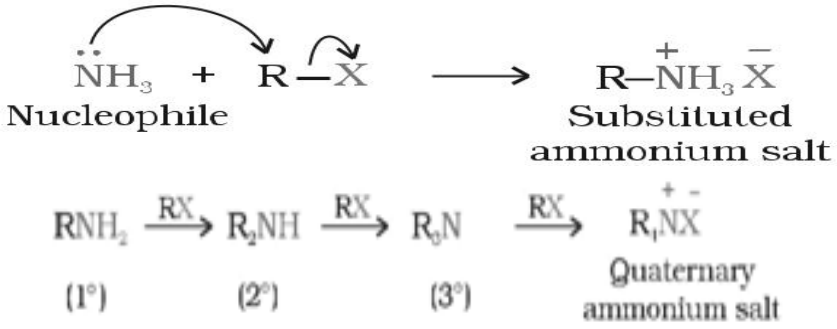
Nitro compounds are reduced to amines by passing hydrogen gas in the presence of finely divided nickel, palladium or platinum and also by reduction with metals in acidic medium. Nitroalkanes can also be similarly reduced to the corresponding alkanamines.



Reduction with iron scrap and hydrochloric acid is preferred because $FeCl_2$ formed gets hydrolysed to release hydrochloric acid during the reaction. Thus, only a small amount of hydrochloric acid is required to initiate the reaction.

2. Ammonolysis of alkyl halides

You have read (Unit 10, Class XII) that the carbon - halogen bond in alkyl or benzyl halides can be easily cleaved by a nucleophile. Hence, an alkyl or benzyl halide on reaction with an ethanolic solution of ammonia undergoes nucleophilic substitution reaction in which the halogen atom is replaced by an amino ($–NH_2$) group. This process of cleavage of the C–X bond by ammonia molecule is known as ammonolysis. The reaction is carried out in a sealed tube at 373 K. The primary amine thus obtained behaves as a nucleophile and can further react with alkyl halide to form secondary and tertiary amines, and finally quaternary ammonium salt.



The free amine can be obtained from the ammonium salt by treatment with a strong base:

$$R–\overset{+}{N}H_3\overset{-}{X} + NaOH \longrightarrow R–NH_2 + H_2O + \overset{+}{N}a\overset{-}{X}$$

Ammonolysis has the disadvantage of yielding a mixture of primary, secondary and tertiary amines and also a quaternary ammonium salt. However, primary amine is obtained as a major product by taking large excess of ammonia.

3. Reduction of nitriles

Nitriles on reduction with lithium aluminium hydride ($LiAlH_4$) or catalytic hydrogenation produce primary amines. This reaction is used for ascent of amine series, i.e., for preparation of amines containing one carbon atom more than the starting amine.

$$R-C\equiv N \xrightarrow[Na(Hg)/C_2H_5OH]{H_2/Ni} R-CH_2-NH_2$$

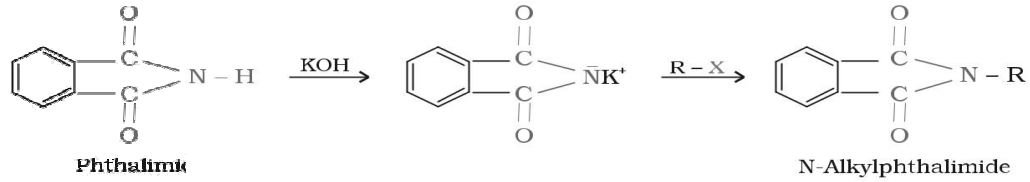
4. Reduction of amides

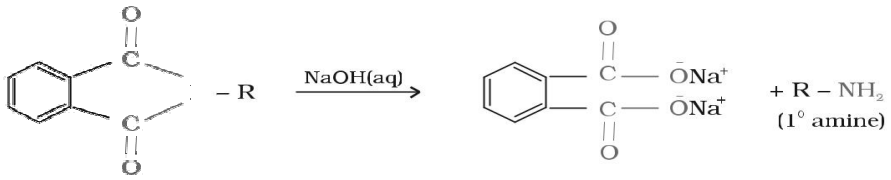
The amides on reduction with lithium aluminium hydride yield amines.

$$R-\overset{\overset{O}{\|}}{C}-NH_2 \xrightarrow[(ii)\ H_2O]{(i)\ LiAlH_4} R-CH_2-NH_2$$

5. Gabriel phthalimide synthesis

Gabriel synthesis is used for the preparation of primary amines. Phthalimide on treatment with ethanolic potassium hydroxide forms potassium salt of phthalimide which on heating with alkyl halide followed by alkaline hydrolysis produces the corresponding primary amine. Aromatic primary amines cannot be prepared by this method because aryl halides do not undergo nucleophilic substitution with the anion formed by phthalimide.





6. Hoffmann bromamide degradation reaction

Hoffmann developed a method for preparation of primary amines by treating an amide with bromine in an aqueous or ethanolic solution of sodium hydroxide. In this degradation reaction, migration of an alkyl or aryl group takes place from carbonyl carbon of the amide to the nitrogen atom. The amine so formed contains one carbon less than that present in the amide.

$$R-\overset{\overset{O}{\|}}{C}-NH_2 + Br_2 + 4NaOH \longrightarrow R-NH_2 + Na_2CO_3 + 2NaBr + 2H_2O$$

1 3 .5 Physical Properties

The lower aliphatic amines are gases with fishy odour. Primary amines with three or more carbon atoms are liquid and still higher ones are solid. Aniline and other arylamines are usually colourless but get coloured on storage due to atmospheric oxidation.

Lower aliphatic amines are soluble in water because they can form hydrogen bonds with water molecules. However, solubility decreases with increase in molar mass of amines due to increase in size of the hydrophobic alkyl part. Higher amines are essentially insoluble in water. Considering the electronegativity of nitrogen of amine and oxygen of alcohol as 3.0 and 3.5 respectively, you can predict the pattern of solubility of amines and alcohols in water. Out of butan-1-ol and butan-1-amine, which will be more soluble in water and why? Amines are soluble in organic solvents like alcohol, ether and benzene. You may remember that

alcohols are more polar than amines and form stronger intermolecular hydrogen bonds than amines.

Primary and secondary amines are engaged in intermolecular association due to hydrogen bonding between nitrogen of one and hydrogen of another molecule. This intermolecular association is more in primary amines than in secondary amines as there are two hydrogen atoms available for hydrogen bond formation in it. Tertiary amines do not have intermolecular association due to the absence of hydrogen atom available for hydrogen bond formation. Therefore, the order of boiling points of isomeric amines is as follows:

Primary > Secondary > Tertiary

Intermolecular hydrogen bonding in primary amines is shown in Fig. 13.2.

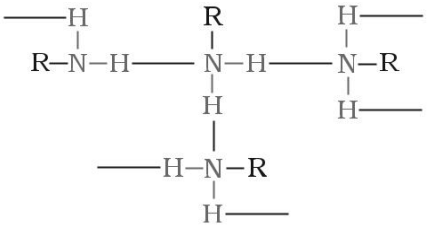


Fig. 13.2 Intermolecular hydrogen bonding in primary amines

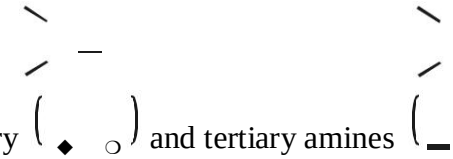
Boiling points of amines, alcohols and alkanes of almost the same molar mass are shown in below Table. 13.2.

Table. 13.2. Comparison of Boiling Points of Amines, Alcohols and Alkanes of Similar Molecular Masses

Sl. No.	Compound	Molar mass	b.p./K
1	n-$C_4H_9NH_2$	73	350.8
2	$(C_2H_5)_2NH$	73	329.3
3	$C_2H_5N(CH_3)_2$	73	310.5
4	$C_2H_5CH(CH_3)_2$	72	300.8
5	n-C_4H_9OH	74	390.3

13.6 Chemical Reactions

Difference in electronegativity between nitrogen and hydrogen atoms and the presence of unshared pair of electrons over the nitrogen atom makes amines reactive. The number of hydrogen atoms attached to nitrogen atom also decides the course of reaction of amines; that is why primary

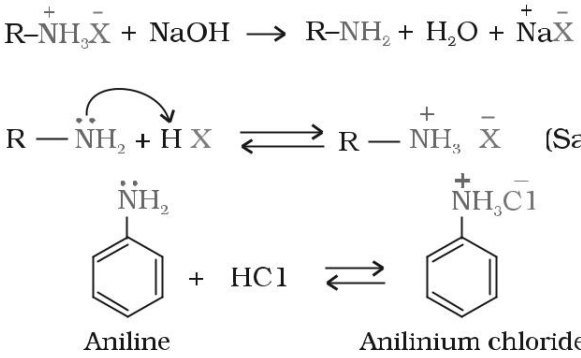


(–NH_2), secondary and tertiary amines differ in many reactions.

Moreover, amines behave as nucleophiles due to the presence of unshared electron pair. Some of the reactions of amines are described below:

Basic character of amines

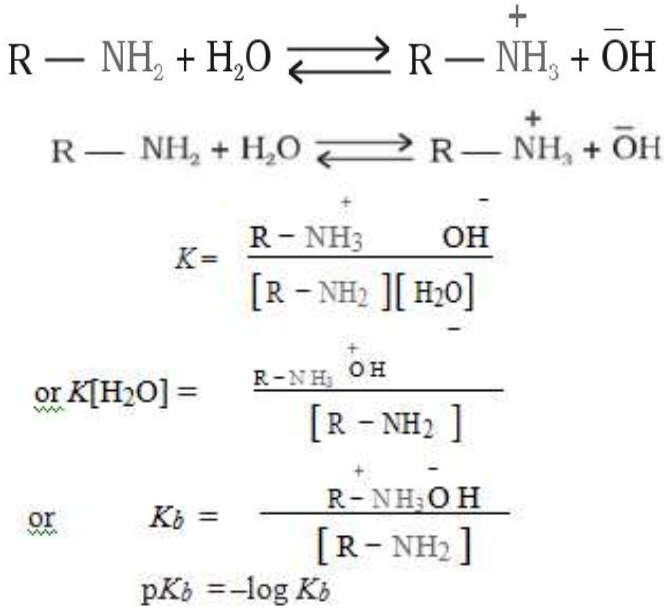
Amines, being basic in nature, react with acids to form salts.



Amine salts on treatment with a base like NaOH, regenerate the parent amine.

Amine salts are soluble in water but insoluble in organic solvents like ether. This reaction is the basis for the separation of amines from the non basic organic compounds insoluble in water.

The reaction of amines with mineral acids to form ammonium salts shows that these are basic in nature. Amines have an unshared pair of electrons on nitrogen atom due to which they behave as Lewis base. Basic character of amines can be better understood in terms of their K_b and pK_b values as explained below:



Larger the value of K_b or smaller the value of pK_b, stronger is the base. The pK_b values of few amines are given in below Table 13.3.

pK_b value of ammonia is 4.75. Aliphatic amines are stronger bases than ammonia due to +I effect of alkyl groups leading to high electron density on the nitrogen atom. Their pK_b values lie in the range of 3 to 4.22. On the other hand, aromatic amines are weaker bases than ammonia due to the electron withdrawing nature of the aryl group.

Table 13.3 pK_b Values of Amines in Aqueous Phase

Name of amine	pK_b
Methanamine	3.38
N-Methylmethanamine	3.27
N,N-Dimethylmethanamine	4.22
Ethanamine	3.29
N-Ethylethanamine	3
N,N-Diethylethanamine	3.25
Benzenamine	9.38
Phenylmethanamine	4.7
N-Methylaniline	9.3
N,N-Dimethylaniline	8.92

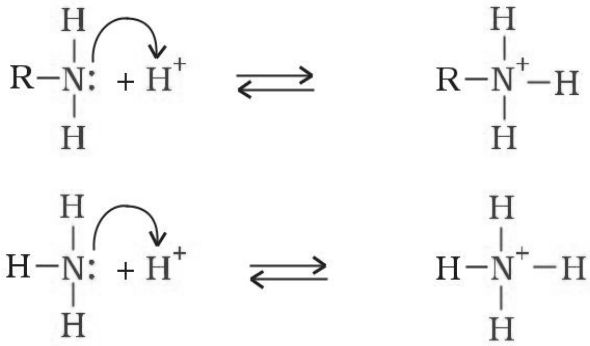
You may find some discrepancies while trying to interpret the K_b values of amines on the basis of +I or –I effect of the substituents present in amines. Besides inductive effect, there are other effects like solvation effect, steric hinderance, etc., which affect the basic strength of amines. Just ponder over. You may get the answer in the following paragraphs.

Structure-basicity relationship of amines

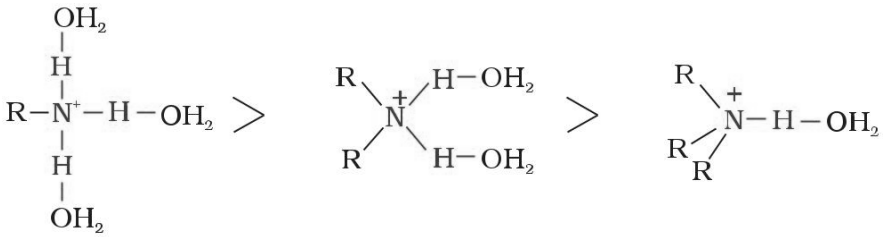
Basicity of amines is related to their structure. Basic character of an amine depends upon the ease of formation of the cation by accepting a proton from the acid. The more stable the cation is relative to the amine, more basic is the amine.

(a) Alkanamines versus ammonia

Let us consider the reaction of an alkanamine and ammonia with a proton to compare their basicity.



Due to the electron releasing nature of alkyl group, it (R) pushes electrons towards nitrogen and thus makes the unshared electron pair more available for sharing with the proton of the acid. Moreover, the substituted ammonium ion formed from the amine gets stabilised due to dispersal of the positive charge by the +I effect of the alkyl group. Hence, alkylamines are stronger bases than ammonia. Thus, the basic nature of aliphatic amines should increase with increase in the number of alkyl groups. This trend is followed in the gaseous phase. The order of basicity of amines in the gaseous phase follows the expected order: tertiary amine > secondary amine > primary amine > NH_3. The trend is not regular in the aqueous state as evident by their pK_b values given in Table 13.3. In the aqueous phase, the substituted ammonium cations get stabilised not only by electron releasing effect of the alkyl group (+I) but also by solvation with water molecules. The greater the size of the ion, lesser will be the solvation and the less stabilised is the ion. The order of stability of ions are as follows:



1° 2° 3°

Decreasing order of extent of H-bonding in water and order of stability of ions by solvation.

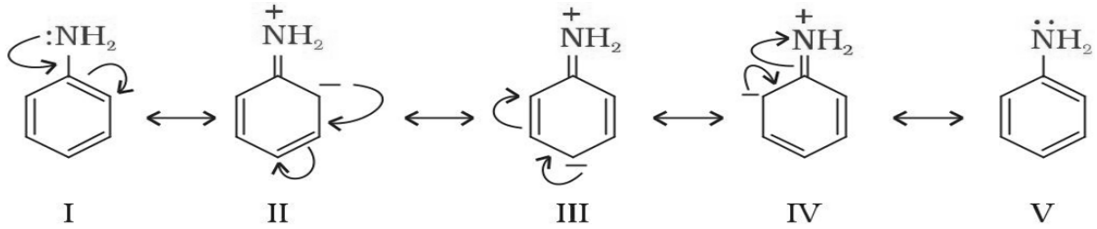
Greater is the stability of the substituted ammonium cation, stronger should be the corresponding amine as a base. Thus, the order of basicity of aliphatic amines should be: primary > secondary > tertiary, which is opposite to the inductive effect based order. Secondly, when the alkyl group is small, like $–CH_3$ group, there is no steric hindrance to H-bonding. In case the alkyl group is bigger than CH_3 group, there will be steric hinderance to H-bonding. Therefore, the change of nature of the alkyl group, e.g., from $–CH_3$ to $–C_2H_5$ results in change of the order of basic strength. Thus, there is a subtle interplay of the inductive effect, solvation effect and steric hinderance of the alkyl group which decides the basic strength of alkyl amines in the aqueous state. The order of basic strength in case of methyl substituted amines and ethyl substituted amines in aqueous solution is as follows:

$$(C_2H_5)_2NH > (C_2H_5)_3N > C_2H_5NH_2 > NH_3$$

$$(CH_3)_2NH > CH_3NH_2 > (CH_3)_3N > NH_3$$

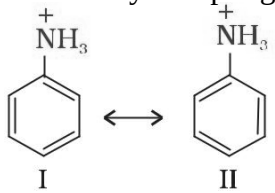
(b) Arylamines versus ammonia

pK_b value of aniline is quite high. Why is it so? It is because in aniline or other arylamines, the $-NH_2$ group is attached directly to the benzene ring. It results in the unshared electron pair on nitrogen atom to be in conjugation with the benzene ring and thus making it less available for protonation. If you write different resonating structures of aniline, you will find that aniline is a resonance hybrid of the following five structures.



I II III IV V

On the other hand, anilinium ion obtained by accepting a proton can have only two resonating structures (kekule).



I II

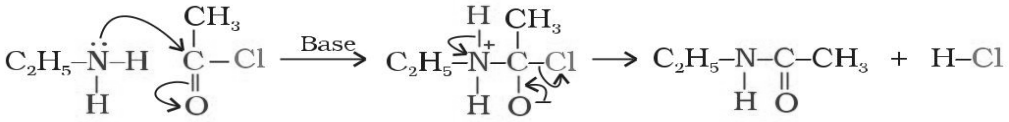
We know that greater the number of resonating structures, greater is the stability. Thus you can infer that aniline (five resonating structures) is more stable than anilinium ion. Hence, the proton acceptability or the basic nature of aniline or other aromatic amines would be less than that of ammonia. In case of substituted aniline, it is observed that electron releasing groups like $–OCH_3$, $–CH_3$ increase basic strength whereas electron withdrawing groups like $–NO_2$, $–SO_3$, –COOH, –X decrease it.

Alkylation

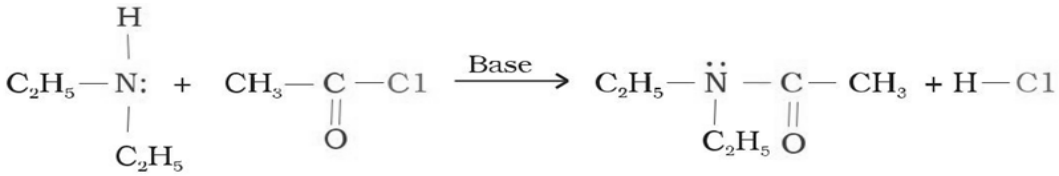
Amines undergo alkylation on reaction with alkyl halides (refer Unit 10, Class XII).

Acylation

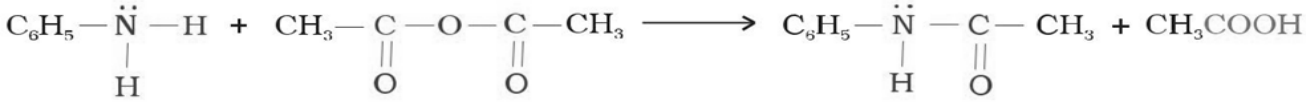
Aliphatic and aromatic primary and secondary amines react with acid chlorides, anhydrides and esters by nucleophilic substitution reaction. This reaction is known as acylation. You can consider this reaction as the replacement of hydrogen atom of $–NH_2$ or >N–H group by the acyl group. The products obtained by acylation reaction are known as amides. The reaction is carried out in the presence of a base stronger than the amine, like pyridine, which removes HCl so formed and shifts the equilibrium to the right hand side.



Ethanamine N-Ethylethanamide



N-Ethylethanamine N,N-Diethylethanamide



Benzenamine Ethanoic anhydride N-Phenylethanamide or Acetanilide

Amines also react with benzoyl chloride (C_6H_5COCl). This reaction is known as benzoylation.

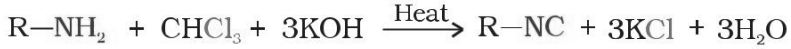
CH_3NH_2 + C_6H_5COCl → $CH_3NHCOC_6H_5$ + HCl

Methanamine Benzoyl chloride N − Methylbenzamide

1. What do you think is the product of the reaction of amines with carboxylic acids ? They form salts with amines at room temperature.

Carbylamine reaction

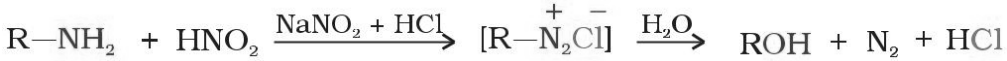
Aliphatic and aromatic primary amines on heating with chloroform and ethanolic potassium hydroxide form isocyanides or carbylamines which are foul smelling substances. Secondary and tertiary amines do not show this reaction. This reaction is known as carbylamine reaction or isocyanide test and is used as a test for primary amines.



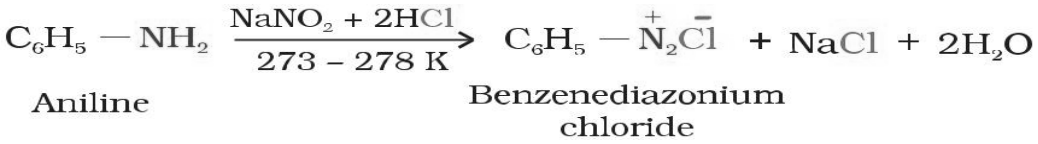
Reaction with nitrous acidyy

Three classes of amines react differently with nitrous acid which is prepared in situ from a mineral acid and sodium nitrite.

(a) Primary aliphatic amines react with nitrous acid to form aliphatic diazonium salts which being unstable, liberate nitrogen gas quantitatively and alcohols. Quantitative evolution of nitrogen is used in estimation of amino acids and proteins.



b) Aromatic amines react with nitrous acid at low temperatures (273-278 K) to form diazonium salts, a very important class of compounds used for synthesis of a variety of aromatic compounds discussed in Section 13.7.

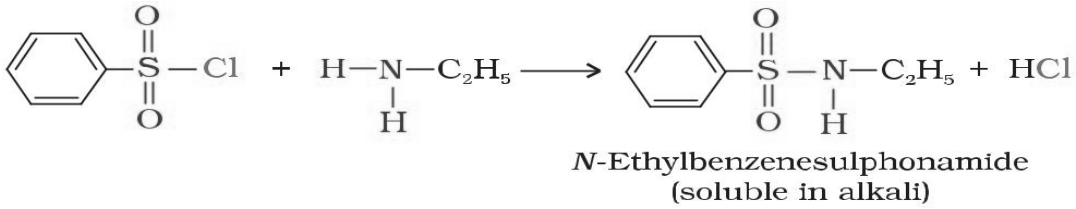


Secondary and tertiary amines react with nitrous acid in a different manner.

Reaction with arylsulphonyl chloride

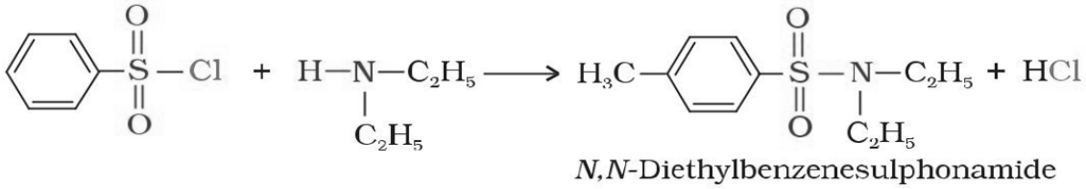
Benzenesulphonyl chloride ($C_6H_5SO_2Cl$), which is also known as Hinsberg's reagent, reacts with primary and secondary amines to form sulphonamides.

The reaction of benzenesulphonyl chloride with primary amine yields N-ethylbenzenesulphonyl amide.



The hydrogen attached to nitrogen in sulphonamide is strongly acidic due to the presence of strong electron withdrawing sulphonyl group. Hence, it is soluble in alkali.

(b) In the reaction with secondary amine, N,N-diethyl-benzenesulphonamide is formed.



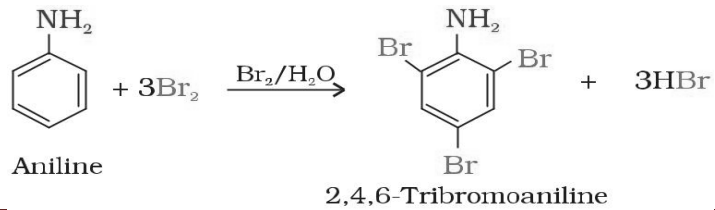
Since N, N-diethylbenzene sulphonamide does not contain any hydrogen atom attached to nitrogen atom, it is not acidic and hence insoluble in alkali.

(c) Tertiary amines do not react with benzenesulphonyl chloride. This property of amines reacting with benzenesulphonyl chloride in a different manner is used for the distinction of primary, secondary and tertiary amines and also for the separation of a mixture of amines. However, these days benzenesulphonyl chloride is replaced by p-toluenesulphonyl chloride.

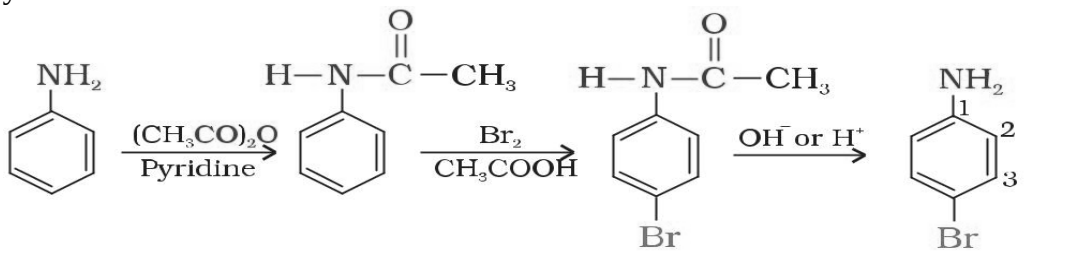
Electrophilic substitution

You have read earlier that aniline is a resonance hybrid of five structures. Where do you find the maximum electron density in these structures? Ortho- and para-positions to the $–NH_2$ group become centres of high electron density. Thus $–NH_2$ group is ortho and para directing and a powerful activating group.

(a) Bromination: Aniline reacts with bromine water at room temperature to give a white precipitate of 2,4,6-tribromoaniline.



The main problem encountered during electrophilic substitution reactions of aromatic amines is that of their very high reactivity. Substitution tends to occur at ortho- and para-positions. If we have to prepare monosubstituted aniline derivative, how can the activating effect of $-NH_2$ group be controlled ? This can be done by protecting the $-NH_2$ group by acetylation with acetic anhydride, then carrying out the desired substitution followed by hydrolysis of the substituted amide to the substituted amine.

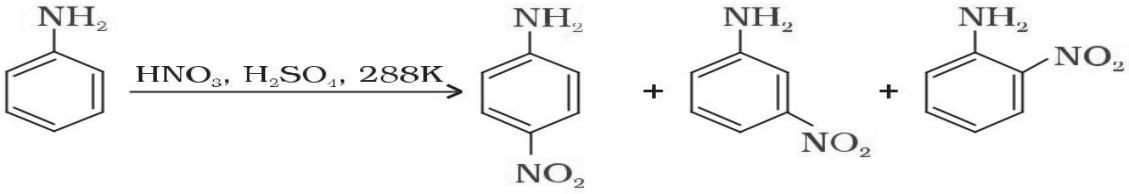


Aniline | *N*-Phenylethanamide (Acetanilide) | (Major) | 4-Bromoaniline

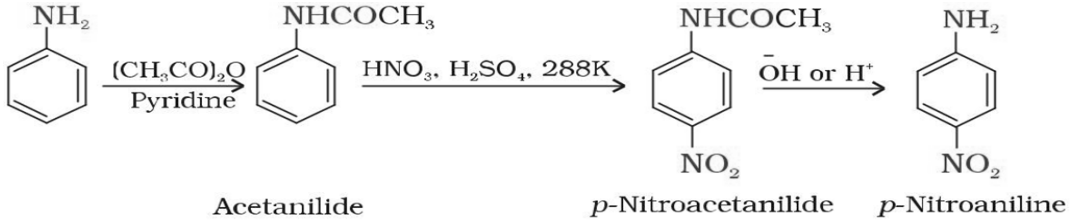
The lone pair of electrons on nitrogen of acetanilide interacts with oxygen atom due to resonance as shown below:

Hence, the lone pair of electrons on nitrogen is less available for donation to benzene ring by resonance. Therefore, activating effect of $-NHCOCH_3$ group is less than that of amino group.

(b) Nitration: Direct nitration of aniline yields tarry oxidation products in addition to the nitro derivatives. Moreover, in the strongly acidic medium, aniline is protonated to form the anilinium ion which is meta directing. That is why besides the ortho and para derivatives, significant amount of meta derivative is also formed.

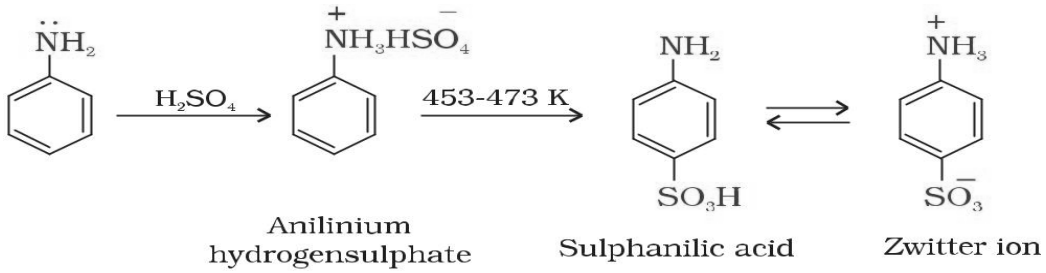


However, by protecting the $-NH_2$ group by acetylation reaction with acetic anhydride, the nitration reaction can be controlled and the p-nitro derivative can be obtained as the major product.



Acetanilide | *p*-Nitroacetanilide | *p*-Nitroaniline

(c) **Sulphonation**: Aniline reacts with concentrated sulphuric acid to form anilinium hydrogensulphate which on heating with sulphuric acid at 453-473K produces p-aminobenzene sulphonic acid, commonly known as sulphanilic acid, as the major product.



Aniline does not undergo Friedel-Crafts reaction (alkylation and acetylation) due to salt formation with aluminium chloride, the Lewis acid, which is used as a catalyst. Due to this, nitrogen of aniline acquires positive charge and hence acts as a strong deactivating group for further reaction.

Summary

Amines can be considered as derivatives of ammonia obtained by replacement of hydrogen atoms with alkyl or aryl groups. Replacement of one hydrogen atom of ammonia gives rise to structure of the type R-NH_2, known as primary amine. Secondary amines are characterised by the structure R_2NH or R-NHR′ and tertiary amines by R_3N, RNR′R″ or R_2NR'. Secondary and tertiary amines are known as simple amines if the alkyl or aryl groups are the same and mixed amines if the groups are different. Like ammonia, all the three types of amines have one unshared electron pair on nitrogen atom due to which they behave as L

Amines are usually formed from nitro compounds, halides, amides, imides, etc. They exhibit hydrogen bonding which influence their physical properties.

In alkylamines, a combination of electron releasing, steric and H-bonding factors influence the stability of the substituted ammonium cations in protic polar solvents and thus affect the basic nature of amines. Alkyl amines are found to be stronger bases than ammonia. In aromatic amines, electron releasing and withdrawing groups, respectively increase and decrease their basic character. Aniline is a weaker base than ammonia. Reactions of amines are governed by availability of the unshared pair of electrons on nitrogen. Influence of the number of hydrogen atoms at nitrogen atom on the type of reactions and nature of products is responsible for identification and distinction between primary, secondary and tertiary amines. p-Toluenesulphonyl chloride is used for the identification of primary, secondary and tertiary amines. Presence of amino group in aromatic ring enhances reactivity of the aromatic amines. Reactivity of aromatic amines can be controlled by acylation process, i.e., by treating with acetyl chloride or acetic anhydride. Tertiary amines like trimethylamine are used as insect attractants.

Aryldiazonium salts, usually obtained from arylamines, undergo replacement of the diazonium group with a variety of nucleophiles to provide advantageous methods for producing aryl halides, cyanides, phenols and arenes by reductive removal of the diazo group. Coupling reaction of aryldiazonium salts with phenols or arylamines give rise to the formation of azo dyes.

IMPORTANT QUESTIONS

1. Write any two Methods of preparations of Amines
2. Write the reactions of Aromatic and Aliphatic primary Amines with Nitrous Acid.
3. Explain the following named reactions
    a. Gabriael Phathalimide synthesis
    b. Hoffmann bromamide degradation reaction

CYANIDES AND ISOCYANIDES

Alkyl cyanides are alkyl derivatives of hydrogen cyanide and isocyanides are isomers of alkyl cyanides.

13.7 Structure of cyanides and isocyanides

A lkyl cyanides are polar compounds and ionize to form alkyl carbo cation (**R^+**) and cyanide ion $^-C \equiv N$**:** The alkyl group in RCN is linked to the carbon of cyanide ion while in isocyanides alkyl group is linked to nitrogen (N) of cyanide.

$R - C \equiv N$ $R\text{-}N \equiv C$

The cyanide ion $^-C \equiv N$: has an unshared pair of electrons on both ends.

Nomenclature:

Alkyl cyanides are named by giving the name of the alkyl group and adding the cyanide. Alkyl cyanides are also called nitriles based on the name of the acid which would be formed upon hydrolysis. For complex compounds cyanides may be considered as cyano derivatives of other classes of compounds.

$CH_3{}^-CN$ Methyl cyanide or acetonitrile.

$CH_3{}^-CH_2{}^-CN$ Ethyl cyanide or propionitrile

Isocyanides (Isonitriles or Carbyl amines) are named as alkyl isocyanides or alkyl isonitriles.

$CH_3{}^-NC$ Methyl isocyanide or Methyl isonitrile

13.8 Preperation

1. From alkyl halides:

Alkyl halides with ethanolic potassium cyanide forms alkyl cyanide as major product while with ethanolic silver cyanide the chief products is the alkyl isocyanide. In both reactions, however some of the other isomer is also produced.

$CH_3{}^-CH_2{}^-Cl \xrightarrow{\text{Ethanolic KCN}} CH_3{}^-CH_2{}^-CN + CH_3{}^-CH_2{}^-NC$

$CH_3{}^-CH_2{}^-Cl \xrightarrow{\text{Ethanolic AgCN}} CH_3{}^-CH_2{}^-CN + CH_3{}^-CH_2{}^-NC$

13.9 Physical Properties:

Cyanides are fairly polar compounds but they are less soluble in water than amines. For example, nitriles higher than propionitrile are only slightly soluble in water. They are much weaker bases than amines too, therefore they are not soluble in aqueous acids.

Alkyl cyanides have pleasant odors and are only slightly toxic, in contrast to isocyanides which have very bad odors and highly toxic. Isocyanides boil at low temperatures than their isomeric cyanides.

13.10 Chemical Properties:

1. Hydrolysis: Cyanides hydrolyze to carboxylic acids and ammonia

$R^-CN + 2H_2O \xrightarrow{H2O^+ \text{ or } ^-OH} RCOOH + NH_3$

Isocyanides on the other hand hydrolyze to form amines and formic acid.

$R^-CN + 2H_2O \xrightarrow{H2O^+ \text{ or } ^-OH} RNH_2 + HCOOH$

2. Reduction: The reduction of nitriles to primary amines

$R^-CN \xrightarrow{R.A} R^-CH_2 - NH_2$ R.A. = $LiAlH_4$, Na, alcohol, H_2/cat.

Reduction of isocyanides yields secondary amines.

$R^-NC \xrightarrow{R.A} R^-NH - CH_3$ R.A. = NaHg, H_2/cat.

BOARD OF INTERMEDIATE EDUCATION- A.P. VIJAYAWADA
VOCATIONAL BRIDGE COURSE
PHYSICS- II YEARS (w.e.f 2019-20)

Weightage of Marks

S.No.	Name of the Chapter	No of periods	Weightage
1.	Waves	05	04
2.	Ray Optics and Optical Instruments	07	04
3.	Wave Optics	03	01
4.	Electric Charges and Fields	04	04
5.	Electrostatic Potential and Capacitance	06	04
6.	Current Electricity	06	04
7.	Moving Charges and Magnetism	06	04
8.	Magnetism and Matter	03	01
9.	Electromagnetic Induction	04	01
10.	Alternating Current	03	01
11.	Electromagnetic Waves	02	01
12.	Dual Nature of Radiation and Matter	04	01
13.	Atoms	02	01
14.	Nuclei	06	04
15.	Semiconductor Electronics: Materials, Devices and Simple Circuits	06	04
16.	Communication Systems	02	01
	Total	69	40

BOARD OF INTERMEDIATE EDUCATION-A.P, VIJAYAWADA
VOCATIONAL BRIDGE COURSE
MODEL QUESTION PAPER (w.e.f 2019-20)

Time: 1 ½ Hours **Paper – II (Physics)** **Max. Marks: 25**

Section – A **5 x 1 = 5**

Instructions:

i) Answer any **five** of the following questions

ii) Each question carries **one** mark.

1. What is diffraction?
2. What are the units of Magnetic Moment and Magnetic Induction?
3. State Lenz Law?
4. What type of transformer is used in 6 V bed lamp?
5. Give one use of infrared rays?
6. Write down Einstein's photoelectric equation?
7. What is the difference between Alpha particle and Helium atom?
8. Which type of communication is employed in mobile phones?

Section – B **5x4=20**

Instructions:

i) Answer any **five** of the following questions

ii) Each question carries **four** marks.

9. Explain the formation of standing waves in a closed pipe.
10. Draw a neat labelled diagram of a simple microscope and explain the formation of the image.
11. State and explain Coulomb's inverse square law in electricity.
12. Three capacitors of capacitance 1μF,2μF and 3μF are connected in parallel.
    (a)What is the ratio of charges? (b)What is the rate of potential difference?
13. Using Kirchoff's laws deduce the condition for balance in a Wheatstone bridge.
14. State and explain Biot-savart law.
15. Distinguish between Nuclear fission and fusion.
16. Explain the working of a full wave rectifier.

BOARD OF INTERMEDIATE EDUCATION, A.P -VIJAYAWADA
VOCATIONAL BRIDGE COURSE
CHEMISTRY – Second Year (w.e.f. 2019-2020)
WEIGHTAGE OF MARKS

S.No.	Name of the Chapter	No of periods	Weightage
1.	Solid State	06	02
2.	Solutions	06	04
3.	Electrochemistry and Chemical Kinetics	10	04
4.	Surface Chemistry	06	04
5.	General Principles of Metallurgy	05	02
6.	p-BLOCK ELEMENTS Group – 15 Elements Group – 16 Elements Group – 17 Elements Group – 18 Elements	16	06
7.	d and f Block Elements, Coordination Compounds	08	04
8.	Polymers	04	02
9.	Bio molecules	04	02
10.	Chemistry in Everyday Life	05	02
11.	Haloakanes and Haloarences	05	02
12.	Organic Compound Containing C, H and O. (Alcohols, Phenols, Ethers, Aldehydes, Ketones and Carboxylic Acids)	10	04
13.	Organic Compounds containing Nitrogen-Amines.	05	02
	Total	90	40

BOARD OF INTERMEDIATE EDUCATION, AP-VIJAYAWADA
VOCATIONAL BRIDGE COURSE
CHEMISTRY – Second Year (w.e.f. 2019-2020)
MODEL QUESTION PAPER

Time: 1 ½ Hours **Max.Marks: 25**

Section – A **5x1=5**

Note:

i) *Answer any **five** of the following questions*

ii) *Each question carries **one** mark.*

1. Define the term unit cell.
2. What is Frankel defect.
3. Explain Polling.
4. What is flux?
5. Draw the structure of XeO_3
6. What is inert pair effect?
7. What is PHBV?
8. Write the name and structure of monomer is Bakelite?

Section – B **5x4=20**

Note:

i) *Answer any **five** of the following questions*

ii) *Each question carries **four** marks.*

9. Define Molarity. Calculate the molarity of a solution containing 5g. of NaOH in 450 ml of solution.
10. Explain Faraday's laws of electrolysis.
11. Write the differences between physical and chemical adsorptions.
12. Describe the manufacture of $H2SO_4$ in contact process.
13. Explain Werner's theory of coordination compounds?
14. (a) Define vitamins (b) What are antibiotics?
15. Write the (a)William sons synthesis(b) Reimer – Tiemann reaction
16. (a) Cannizaro reaction (b) De-carboxylation.